Un approccio basato su DBpedia per la sistematizzazione della conoscenza sul Web

POLITECNICO DI TORINO

SCUOLA DI DOTTORATO
Dottorato in Beni Culturali - XXV Ciclo

Tesi di Dottorato

Un approccio basato su DBpedia per la sistematizzazione della conoscenza sul Web

Federico Cairo

Tutore
prof. Mario Ricciardi

Coordinatore del corso di dottorato
prof. Costanza Roggero

Marzo 2013

Indice

Ringraziamenti

Desidero innanzitutto ringraziare il Prof. Mario Ricciardi, tutore del mio percorso di dottorato, per il sostegno e l'aiuto fornitomi nella scelta e nello sviluppo del tema di questa tesi. La mia gratitudine va inoltre al Centro Nexa su Internet & Società del Politecnico di Torino, dove ho svolto buona parte del mio lavoro su TellMeFirst e dove ho avuto occasione di approfondire le tematiche relative ai Linked Open Data in un ambiente di grande competenza e cordialità. Voglio anche esprimere la mia gratitudine nei confronti di Telecom Italia (in particolare Walter Goix, Fabio Mondin, Carmen Crimisi e Carlo Alberto Licciardi) per aver creduto nel progetto TellMeFirst tanto da finanziarlo e seguirne lo sviluppo in tutte le sue fasi. Grazie a CSI Piemonte e a CELI (in particolare Giuliana Bonello, Nathalie Coué, Andrea Muraca, Alessandro Trombotto e Vittorio Di Tomaso) per avermi formato "sul campo" nei primi tempi della mia esperienza dottorale. Un ultimo grande ringraziamento va a Giuseppe Futia e a Federico Benedetto, sviluppatori del front-end di TellMeFirst, brillanti collaboratori e soprattutto amici, senza i quali non ce l'avrei mai fatta.

Introduzione

Nel 2004, quando ho scritto la mia tesi di laurea in Filosofia all'Università di Bologna, il Web 2.0 cominciava a muovere i primi passi incerti e il compito di un tesista era più semplice. Si raccoglievano i testi più rilevanti su di un argomento, per lo più cartacei, in base alle bibliografie contenute in altri testi cartacei sullo stesso tema. Inizialmente i titoli sconosciuti erano parecchi, ma ben presto cominciavano a ripetersi, le novità diminuivano di numero e si giungeva alla piacevole situazione in cui i documenti citati erano sempre gli stessi: in quel momento ci si sentiva di padroneggiare una materia, per quanto piccola e insignificante rispetto alla totalità della conoscenza umana.

Chi deve scrivere una tesi oggi, invece, in un'epoca di maggiore maturità del Web, si trova nella complicata posizione di dover porre dei limiti estremamente soggettivi alle proprie letture. Il rapporto numerico tra i testi digitali reperibili online e quelli pubblicati su carta, soprattutto in merito a un argomento di recente interesse, è incredibilmente alto, ed è in costante crescita. Al 3 ottobre 2012, digitando "Semantic Web" sul motore di ricerca del Catalogo del Servizio Bibliotecario Nazionale[1], otteniamo 98 risultati, ma digitando le stesse *keyword* su una nota piattaforma di *social sharing* per paper scientifici, Mendeley[2], il numero sale fino a 8.770 (circa 90 volte maggiore). I risultati della stessa ricerca su Google, si assestano invece sui 8.840.000 (un numero circa 90.000 volte maggiore). Si tratta di un muro quasi minaccioso di documenti, una situazione che fa subito venire in mente parole come sovraccarico informativo o «infobesity» (Johnson, 2012).

Siamo arrivati al punto in cui la quantità di informazioni reperibili su Internet ci impedisce di fatto di conoscere un argomento? Alcuni sono di quest'avviso. Nicolas Carr per esempio, nel suo famoso saggio *The shallows: what the Internet is doing to our brains*, mette in guardia sulla facilità di ottenere a portata di mano enormi quantità di dati e informazioni senza la minima capacità

[1] URL: http://www.sbn.it/opacsbn/opac/iccu/free.jsp

[2] URL: http://www.mendeley.com/

di saperli discernere (Carr, 2010). Oppure Umberto Eco, che sempre più spesso in interviste e articoli manifesta il suo pessimismo nei confronti del Web come strumento conoscitivo, a causa della «censura per eccesso di rumore» (Eco, 2012) che caratterizza l'*information overload*[3].

Combattere il sovraccarico informativo è un tema centrale di questa tesi. Concordando con la prospettiva delineata da David Weinberger nel suo recente saggio *Too Big To Know*, la strategia più efficace e meno costosa per affrontare l'eccesso di conoscenza disponibile oggi è quella di aggiungere informazione all'informazione, sotto forma di metadati (Weinberger, 2011). I Linked Open Data costituiscono un deposito estremamente ricco ed interconnesso di metadati, che sono identificati in maniera univoca attraverso gli URI (Uniform Resource Identifier), liberando il campo da ambiguità e fraintendimenti. Le tecnologie del Web Semantico in generale, e DBpedia in particolare, facilitano l'inserimento di informazioni aggiuntive sul significato dei documenti in Internet e sulla loro fonte, attribuita attraverso URI nel dominio di enti certificati. Questo costituisce sicuramente un passaggio importante per migliorare la nuova infrastruttura della conoscenza veicolata dal Web.

Nella tesi si propone una strategia per la sistematizzazione della conoscenza sul Web basata sui concetti presenti in DBpedia e finalizzata alla riduzione dell'*information overload*. L'opportunità dell'utilizzo di DBpedia è sostenuta sia da aspetti puramente tecnici sia da valutazioni più teoriche. Essendo DBpedia collegata a un vasto corpus multilingue preannotato di carattere enciclopedico (Wikipedia), essa risulta tecnicamente molto adatta ad essere utilizzata per procedimenti automatici di Natural Language Processing e di Text Mining. Ma la cosa più interessante è che i concetti presenti in DBpedia sono il risultato di un consenso semantico raggiunto in maniera collaborativa dalla comunità degli internauti. Un criterio efficace di classificazione sul Web, infatti, non può essere imposto dall'alto, ma deve seguire gli stessi principi di libertà e trasparenza che hanno da sempre costituito l'essenza di Internet.

[3] Per una trattazione meno informale dell'argomento da parte di Eco, si veda l'intervista rilasciata a Patrick Coppock *A Conversation on Information* (Eco, 1995).

A rischio di creare un meccanismo autoreferenziale, devono essere gli utenti stessi di Internet a decidere sia le classi in cui categorizzare i documenti sia il significato che si attribuisce a tali classi. La *knowledge base* utilizzata come ontologia su Internet dev'essere la conoscenza condivisa dagli utenti, giusta o sbagliata che sia. Wikipedia serve perfettamente a questo scopo, in quanto rappresenta il cervello collettivo e continuamente mutante della Rete.

Il primo capitolo della tesi descrive Wikipedia come un frutto di quell'intelligenza collettiva e di quella cultura collaborativa che sembrano emergere come i tratti costitutivi delle comunità in Rete. Vengono esaminate le posizioni di diversi autori sui concetti di intelligenza collettiva (come Pierre Lévy, James Surowiecki, David Weinberger, Micheal Nielsen) e di cultura collaborativa (tra cui Yochai Benkler, Manuel Castells, Henry Jenkins, Eric Raymond, Raffaele Meo). È proposta un'analisi dei punti di forza e di debolezza di Wikipedia per cercare di capire come tali aspetti possano influenzare la sua validità quale corpus annotato per la classificazione dei documenti online.

Il secondo capitolo prende in esame DBpedia, inserendola nel contesto più ampio dei Linked Open Data. Si focalizza sui meccanismi tecnici che permettono la trasformazione della conoscenza non strutturata presente in Wikipedia nella conoscenza strutturata di DBpedia. DBpedia è vista come lo strumento più adatto per costruire un'ontologia della Rete condivisa e distribuita e per sistematizzare la conoscenza presente su Internet.

Nel terzo capitolo viene descritta una soluzione software basata sull'utilizzo di tecnologie semantiche in grado di classificare automaticamente i documenti sul Web sulla base delle risorse presenti in DBpedia. Sono esposti il funzionamento e la metodologia del software TellMeFirst[4], sviluppato dal tesista nell'arco del 2011-2012 all'interno del Dipartimento di Automatica ed Informatica del Politecnico di Torino e in virtù di un *grant* Working Capital da parte di Telecom Italia.

Nel quarto capitolo è delineato un possibile scenario futuro, frutto di questo processo di classificazione. Ogni concetto presente in Wikipedia diventa un Gateway per un insieme di documenti ordinati secondo la loro attinenza all'argomento stesso. Ognuno di questi Gateway si può configurare come un mo-

[4] URL: http://tellmefirst.polito.it/

tore di ricerca semantico su un sottoinsieme di documenti del Web, dove si possono effettuare ricerche specifiche per sottoargomento o per argomenti correlati. I DBpedia Gateways possono essere contenuti o linkati direttamente nelle pagine di Wikipedia, come punto di partenza per approfondire un argomento specifico.

Capitolo 1

Wikipedia e l'intelligenza collettiva

1.1 Il successo di Wikipedia

Dati alla mano, Wikipedia è una delle scommesse più riuscite della storia di Internet. Nata nel 2001 col nome di Nupedia, un progetto di enciclopedia digitale che doveva competere con le edizioni online di Microsoft Encarta e Britannica, già nei primi 12 mesi di vita ha raggiunto i 20000 articoli in 18 lingue (Lih, 2009, p. 104). Nel 2005 Wikipedia è diventata, secondo un rapporto del sito Hitwise[5], il portale di consultazione (*reference site*) più popolare del mondo. Secondo l'ultimo rilevamento di Alexa[6] (11 settembre 2012) Wikipedia è posizionato al sesto posto nella classifica dei siti più visitati del Web, dopo Google, Facebook, YouTube, Yahoo e Baidu. Oggi le statistiche relative alla diffusione di Wikipedia variano a una tale velocità da diventare subito obsolete, tanto che la Wikimedia Foundation[7] fornisce meccanismi automatici di aggiornamento dei dati statistici in pagine apposite come "Wikipedia:Statistics"[8] e "Special:Statistics"[9]. Questi dati restituiscono l'impressionante cifra di 4.094.266 articoli al 10 novembre 2012 per la sola Wikipedia in lingua inglese[10] (23.682.568 articoli in totale), con più di 37 milioni di utenti registrati in 275 lingue[11].

Anche a prescindere dai numeri, il successo di Wikipedia è facilmente riconoscibile nella nostra vita quotidiana. Essendo sempre tra i primi risultati delle ri-

[5] URL: http://www.hitwise.com/uk
[6] URL: http://www.alexa.com/siteinfo/wikipedia.org
[7] URL: http://wikimediafoundation.org/wiki/Home
[8] URL: http://en.wikipedia.org/wiki/Wikipedia:Statistics
[9] URL: http://en.wikipedia.org/wiki/Special:Statistics
[10] URL: http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia
[11] URL: http://meta.wikimedia.org/wiki/List_of_Wikipedias#Grand_Total

cerche su Google e su altri motori, Wikipedia è per la maggior parte di noi lo strumento principale per controllare informazioni al volo, per chiarire dubbi o anche per approfondire particolari argomenti, tanto che Laurence Lessig di recente ha osservato: «Now, none of us understands anything new without first pinging Wikipedia's brain to see its cut on whatever piques our curiosity» (Reagle Jr, 2010, p. IX). Per quanto Wikipedia sia ancora vista con scetticismo o apertamente osteggiata da una parte del mondo scolastico e accademico, l'enciclopedia sta diventando una fonte di riferimento per intere generazioni di studenti. In particolar modo negli argomenti scientifici e tecnologici in lingua inglese, Wikipedia è largamente apprezzata come fonte affidabile e aggiornata. Come scrive Cass Sunstein in *Infotopia* (Sunstein, 2006, p. 154), «In areas that involve technology, Wikipedia tends to shine, often outperforming ordinary encyclopedias – a tribute to the technology-savvy participants that it attracts». Negli ultimi anni Wikipedia ha destato sempre più interesse in ambito accademico, sia come fenomeno sociale da indagare sia come risorsa per lo studio e l'apprendimento. Sono sempre più numerose le ricerche internazionali dedicate specificamente a Wikipedia, raccolte meticolosamente da Wikipedia stessa in pagine come "Academic studies of Wikipedia"[12] o "Wikipedia in research"[13].

Inoltre Wikipedia è uno dei siti più linkati della Rete: un'enorme quantità di pagine esterne produce costantemente collegamenti a pagine dell'enciclopedia. Secondo le statistiche di SEOprofiler[14], Wikipedia è posizionata al settimo posto nella classifica dei siti che hanno un maggior numero di *backlink*, ovvero di link in ingresso. Ma i collegamenti ipertestuali verso Wikipedia hanno un carattere diverso rispetto a quelli dei principali concorrenti in questa classifica (Twitter, Facebook, Youtube, Wordpress, Google). Non sono link a contenuti originali, introvabili altrove, come un video, un *twit* o un profilo Facebook. L'utente che crea un link verso Wikipedia, sceglie quella particolare fonte tra tutte le altre che trattano lo stesso argomento in Rete, come siti istituzionali, divulgativi, blog o altre enciclopedie. I link verso Wikipedia vengono creati sia per spostare altrove

[12] URL: http://en.wikipedia.org/wiki/Wikipedia:Academic_studies_of_Wikipedia
[13] URL: http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_research
[14] URL: http://www.seoprofiler.com/statistics/top-backlinks/world

l'approfondimento di un concetto, sia per rendere tale concetto "univoco", non ambiguo semanticamente. Da questo punto di vista, di fatto Wikipedia nel tempo ha vinto una sorta di selezione naturale che l'ha portata ad essere usata come ontologia di riferimento sul Web. Per quanto scarne, superficiali o incomplete possano essere le sue voci, sono il frutto di un "accordo semantico" degli utenti di Internet che attribuisce ad ogni voce un concetto univoco. È questo l'aspetto di Wikipedia maggiormente interessante nell'ambito di questa tesi. Qui non si intende dimostrare che Wikipedia è una risorsa di consultazione affidabile, né che le sue voci possono competere in accuratezza o estensione a quelle dell'Enciclopedia Britannica[15] e della Treccani[16]. Ciò che maggiormente importa è invece l'affermazione di Wikipedia sul Web quale principale riferimento concettuale condiviso tra gli utenti di Internet. Il fatto che Wikipedia abbia 23 milioni di voci in 275 lingue e che ognuna di queste voci sia connessa con un equivalente nelle altre lingue, significa che gli internauti di tutto il mondo possono concordare sul significato di 23 milioni di parole. Gli aspetti che riguardano la disambiguazione delle voci sono tenuti in grande conto dalla comunità di Wikipedia, perché sono uno dei presupposti principali per il suo funzionamento.

> Disambiguation in Wikipedia is the process of resolving conflicts in article titles that occurs when a single title could be associated with more than one article. In other words, disambiguations are paths leading to different articles which could, in principle, have the same title. (Reagle Jr, 2010, p. 98)

Ancor più importante è il fatto che la semantica di queste voci non è decisa dall'alto, ma nasce da uno spontaneo movimento *grassroots*. Sono gli utenti stessi di Internet che concordano tra loro i significati dei concetti che utilizzano su Internet. Da questo punto di vista, l'aderenza di Wikipedia col mondo reale è di secondaria importanza: l'universo della Rete potrebbe anche essere completamente autoreferenziale, ma Wikipedia sarebbe in ogni caso lo strumento più adatto per la sistematizzazione della conoscenza che vi è contenuta.

[15] URL: http://www.britannica.com/
[16] URL: http://www.treccani.it/

La riuscita del progetto Wikipedia è dovuta in gran parte alla sua policy, ovvero all'insieme di norme che la comunità dei wikipediani riconosce come costitutive della propria identità. Le tre norme fondamentali sono state inizialmente elaborate dal professore di filosofia Larry Sanger, co-fondatore di Wikipedia insieme a Wimmy Wales. In seguito hanno subito vari aggiustamenti, aggiunte ed elaborazioni da parte dell'intera comunità. Esse sono:

1) Neutral Point of View;
2) Verifiability;
3) No Original Research.

Il principio del Neutral Point of View consiste nel semplice assunto che un'enciclopedia non dovrebbe privilegiare una specifica interpretazione della materia trattata rispetto a un'altra, né manifestare l'influenza di particolari bias o pregiudizi. Tutti i punti di vista su un particolare argomento devono avere lo stesso valore e dunque lo stesso peso in termini di trattazione. Questa prospettiva non implica la creazione di voci del tutto prive di giudizi, ma la compresenza nelle singole voci di tutto il ventaglio dei diversi giudizi.

> Wales acknowledged the impossibility of being truly neutral or objective, but he contended , «One of the great things about NPOV is that it is a term of art, and a community fills it with meaning over time». While it may be impossible to achieve true neutrality, the policy has worked remarkably well. The community has rallied around the idealistic vision of coming up with a single unified treatment of any given topic. (Lih, 2009, p. 9)

Dunque, benché Wikipedia abbia una sola pagina per ogni argomento, in ognuna di queste pagine dovrebbero essere riportati i giudizi opposti e le controversie in quell'ambito. Tale visione rispecchia da vicino la deontologia tipica del giornalismo di matrice anglosassone, dove, idealmente, prima vengono i fatti accertabili e poi le opinioni e queste ultime sono considerate tutte aventi peso equivalente. Che non si tratti di una semplice affermazione di principio, ma di una regola realmente applicata, è testimoniato dal giudizio critico dello storico americano Roy Rosenzweig, per cui le voci di Wikipedia spesso risultano una noiosa sequela di fatti e di opinioni contrapposte, senza mai un giudizio profondo o

un'analisi illuminante (Rosenzweig, 2006)[17]. Wikipedia rinuncia spesso alla profondità per non infrangere la norma del Neutral Point of View: quando all'interno della comunità scoppia una "guerra di revisione" dovuta a giudizi contrapposti su uno stesso oggetto, la soluzione più adottata è quella di trovare una formulazione che mantenga entrambi i punti di vista e ne evidenzi la contrapposizione.

> While NPOV policy at first seems like an impossible, or even naïve, reach toward an objectively neutral knowledge, it is quite the opposite. The NPOV policy instead recognizes the multitude of viewpoints and provides an epistemic stance in which they all can be recognized as instances of human knowledge – right or wrong. (Reagle Jr, 2010, p. 11)

Il secondo principio, la verificabilità, ha in comune con il primo la centralità attribuita ai fatti. Il termine *fact* è tradotto in italiano con "fatto", ma è usato molto più comunemente nei paesi anglosassoni rispetto a quanto non avvenga in Italia. In particolare nel giornalismo, nella politica e nella giurisprudenza, *fact* esprime un dato specifico verificabile e la cui fonte sia attribuibile. L'ossessione americana per i fatti è stata analizzata molto bene da David Weinberger nel suo libro *Too Big To Know* (Weinberger, 2011, pp. 19-41). Senza un fatto concreto non può esistere una scoperta, e senza una scoperta non può esistere una qualche forma di conoscenza. Jimmy Wales, il padre di Wikipedia, è sempre stato un sostenitore della filosofia oggettivista di Ayn Rand, e parte della concezione della conoscenza presente nel pensiero della Rand si ritrova in molti principi sostenuti da Wales e Sanger (Lih, 2009, p. 39). Secondo l'oggettivismo, vi è una realtà di oggetti che non dipende dalla mente di ogni singolo individuo. Compito di un'enciclopedia è raccogliere questi fatti e fornire una chiara attribuzione della fonte da cui sono ottenuti. La verificabilità della fonte si rivela comunque un problema abbastanza complesso quando si ha a che fare con milioni di contributori che citano milioni di fonti. Sarebbe impossibile per la comunità dei wikipediani verificare ogni singola citazione di opere cartacee, che possono essere volumi situati in luoghi remoti e difficilmente accessibili. Per questo, sebbene non vi sia

[17] Si veda il paragrafo 1.3.

una regola scritta, la comunità privilegia *de facto* le fonti digitali liberamente accessibili sul Web, la cui verifica è semplice e immediata. I critici hanno rilevato in questa caratteristica una "autoreferenzialità digitale" di Wikipedia, un circolo vizioso che impedisce a Wikipedia di arricchirsi al di fuori del Web[18]. Tuttavia, se pensiamo che il processo di digitalizzazione delle opere cartacee procede a un ritmo sostenuto, di pari passo coi movimenti Open Access[19] e Open Content[20], possiamo ritenere che in un prossimo futuro questo problema sarà meno grave di quanto non sia oggi e che in generale tenda ad assottigliarsi nel tempo. In relazione all'editoria, infatti, mondo cartaceo e mondo digitale stanno sempre più confondendo le loro delimitazioni per convergere verso la Rete.

Il terzo principio, No Original Research, costituisce un'ulteriore specificazione del principio di Verifiability. Non soltanto le affermazioni contenute in Wikipedia devono avere una fonte verificabile, ma tale fonte deve essere preferibilmente «reliable and independent of the subject» (Lih, 2009, p. 137). Non c'è spazio per la creatività o per la scoperta di nuovi orizzonti del pensiero: citare come fonte il proprio blog personale o uno sconosciuto gruppo di opinione, viola questa norma della policy. Sono invece preferiti i contribuiti che attingono da risorse considerate affidabili dalla maggior parte dell'opinione pubblica (occidentale) come paper scientifici, articoli di testate giornalistiche importanti, riviste specializzate, media istituzionali.

Insieme ai tre principi appena esposti, ve n'è un quarto che può essere considerato piuttosto una *best practice*, ovvero un consiglio che la comunità dei wikipediani da ai nuovi arrivati: Assume Good Faith, presumi la buonafede. Questo è particolarmente importante perché si collega con la tematica della cultura collaborativa che verrà trattata nel paragrafo successivo. La creazione e il mantenimento di una voce di Wikipedia è un processo molto meno lineare di quel che si crede. Accedendo alle sezioni "Talk" e "View history" di un articolo, ci si rende immediatamente conto della quantità di revisioni, commenti, discussioni, talvolta addirittura guerre, che si nascondono dietro la sua placida facciata. Il metodo

[18] Si veda par. 1.3.

[19] URL: http://oad.simmons.edu/oadwiki/Main_Page

[20] URL: http://www.opencontent.org/

con cui i wikipediani scrivono l'enciclopedia è basato sulla discussione, sullo scambio ragionato di opinioni, perciò in un contesto simile l'assunzione della buonafede del proprio interlocutore è vitale. Altrimenti i conflitti scoppierebbero ogni piè sospinto, per colpa di accuse, recriminazioni e sospetti.

> Unless there is strong evidence to the contrary, assume that people who work on the project are trying to help it, not hurt it; avoid accusing others of harmful motives without particularly strong evidence. (Lih, 2009, p. 134)

Wikipedia sostiene esplicitamente gli ideali di collaborazione e di tolleranza tra gli esseri umani, anzi vuole mostrarsi al mondo come una prova che questi ideali possono trovare un'applicazione concreta nella realtà. In un intervista rilasciata alla CNN nel dicembre del 2005, Jimmy Wales ha sottolineato come, nonostante tutti i suoi difetti e la sua immaturità, è innegabile che Wikipedia abbia dimostrato che non c'è motivo per presupporre che gli altri siano sempre peggiori, meno razionali e più ottusi di noi: il sogno della collaborazione in buonafede è possibile, non è una semplice utopia.

> Generally we find most people out there on the Internet are good, I mean that's one of the wonderful humanitarian discoveries on Wikipedia is that most people only want to help us build this free, non-profit charitable resource.[21]

Gli ideali collaborativi presenti in Wikipedia sono gli stessi della cultura FOSS (Free and Open Source Software). Wales era molto affascinato dall'etica hacker e dal movimento *open source*, tanto da entrare personalmente in contatto col padre fondatore del *copyleft* Richard Stalman, il quale nel 2000 gli consigliò di adottare la GNU Free Documentation License[22] per il neonato progetto Nupedia. Lo stesso Stalman aveva avuto qualche anno prima l'idea di una enciclopedia libera e *open source*, che si ispirasse alla filosofia del sistema operativo GNU, a cui voleva dar nome Gnupedia. In seguito alla na-

[21] La trascrizione dell'intervista si trova all'URL:
http://en.wikipedia.org/wiki/User:One/Wales_interview_transcript
[22] URL: http://www.gnu.org/copyleft/fdl.html

scita di Nupedia, Stalman decide di far convergere Gnupedia in Nupedia, per non creare inutili competizioni tra due progetti identici[23].

Ciò che più accomuna Wikipedia alla filosofia FOSS è perfettamente sintetizzato dalla celebre frase di Eric Raymond, l'autore del saggio *La cattedrale e il bazaar*: «Given enough eyeballs, all bugs are shallow» (dato un numero sufficiente di occhi, tutti i bug vengono a galla), affermazione nota anche come "Legge di Linus", dal nome di Linus Torvalds, creatore del kernel Linux.

Questi aspetti sono trattati efficacemente da Sara Monaci nel breve saggio *La conoscenza on line: logiche e strumenti* (2008). La Monaci individua due tendenze fondamentali all'interno della Internet odierna, due matrici opposte che hanno trovato un equilibrio dinamico: quella tecnocratica incarnata da Google e quella comunitaria esemplificata da Wikipedia.

> Nell'attuale Word Wide Web convivono culture e modelli prettamente tecnocratici con le sperimentazioni sociali di costruzione e codifica delle conoscenze fondate sulla trasparenza, sulla collaborazione fra pari e su un'etica ideal-comunitaria. (Monaci, 2008, p. 11)

Questi modelli contrapposti influenzano il nostro modo di utilizzare Internet in relazione all'accesso, alla creazione e alla condivisione della conoscenza. Mentre l'«interfaccia mondo» di Google offusca tutta la complessità della Rete dietro alla semplice azione della ricerca, Wikipedia mostra, svela i meccanismi di produzione della conoscenza, la loro complessità e instabilità. Una delle maggiori novità di Wikipedia rispetto alle enciclopedia tradizionali è la totale trasparenza del processo, tanto che il valore del risultato è pari a quella della sua costruzione. Mentre l'algoritmo di Google impoverisce l'utente della sua capacità di costruire significati, il modello di Wikipedia arricchisce l'utilizzatore grazie al continuo scambio concettuale finalizzato al consenso, alla mediazione attraverso il dialogo. Scambio che non si esaurisce nel momento in cui viene vissuto, ma che rimane scritto e serve da "storia" della comunità, da esempio comportamentale per i nuovi arrivati.

[23] Si veda il manifesto di Gnupedia all'URL:
http://www.gnu.org/encyclopedia/encyclopedia.en.html

1.2 Intelligenza collettiva e cultura partecipativa in Wikipedia

I concetti di cultura collaborativa e di intelligenza collettiva sono strettamente correlati ed aiutano a comprendere la complessità del fenomeno Wikipedia. Alcuni autori, come Pierre Lévy (1994), Clay Shirky (2008), David Weinberger (2011) e Michael Nielsen (2012) tendono a trattare questi argomenti come un tutt'uno, per evidenziare come nella Rete la collaborazione *networked* sia in grado di far emergere l'intelligenza dalle masse. In effetti il successo di Wikipedia ha a che vedere con l'attivazione dei meccanismi dell'intelligenza collettiva che è consentita proprio dalla collaborazione. Per questo Wikipedia può essere considerata come una prova *a posteriori* del fatto che l'intelligenza collettiva e la cultura collaborativa agiscono efficacemente nella realtà del Web (Benkler, 2002; Jenkins, 2006; Shirky, 2008).

La nozione di intelligenza collettiva nasce nell'ambito della biologia (più esattamente dell'etologia animale) per descrivere il comportamento di alcune specie di insetti (come le formiche e le api) che agiscono in gruppo come un unico superorganismo. Ad utilizzare per la prima volta questa espressione è stato l'entomologo americano William Morton Wheeler, che nell'articolo del 1911 *The ant-colony as an organism* ha descritto le colonie di formiche come unità sociali in cui diversi agenti individuali con limitata intelligenza e informazione erano in grado di riunire le risorse per raggiungere obiettivi fuori dalla portata delle capacità individuali (Wheeler, 1911).

Benché alcune concettualizzazioni vicine all'idea di intelligenza collettiva si possano ravvisare anche in Durkeim (1912)[24], è stato il sociologo americano Ho-

[24] Émile Durkheim in *Les formes élémentaires de la vie religieuse* esprime l'idea secondo cui la società rappresenta l'intelligenza più elevata perché trascende l'individuo nello spazio e nel tempo: «Ci si meraviglierà forse di vederci riferire alla società le forme più elevate della mentalità umana. […] Se le sintesi delle rappresentazioni particolari che si producono in seno a ogni coscienza individuale sono già, da sole, produttrici di novità, quanto più efficaci saranno queste vaste sintesi di coscienze complete che sono le società! La società è il più fiorente fascio di forze fisiche e morali di cui la natura ci offra lo spettacolo. In nessun luogo si trova una tale ricchezza di materiali diversi, portata a un simile grado di concentrazione. Non è quindi sorprendente

ward Bloom il primo a trasferire questo concetto dal campo biologico a quello delle scienze umane, individuando una linea di continuità che va dal comportamento collettivo dalle piante agli animali fino alla società umana. Bloom ha mostrato come l'intelligenza collettiva delle colonie di batteri e delle società umane siano entrambe replicabili attraverso sistemi adattativi complessi («complex adaptive systems») generati dal computer e governati da algoritmi genetici (Bloom, 1995).

L'espressione "intelligence collective" ha trovato maggiore diffusione in Europa grazie al filosofo francese Pierre Lévy, che nel saggio intitolato *L'intelligence collective. Pour une anthropologie du cyberespace* (1994), l'ha così definita:

> È un'intelligenza distribuita ovunque, continuamente valorizzata, coordinata in tempo reale, che porta a una mobilitazione effettiva delle competenze. Aggiungiamo alla nostra definizione questa precisazione indispensabile: il fondamento e il fine dell'intelligenza collettiva sono il riconoscimento e l'arricchimento reciproco delle persone, e non il culto di società feticizzate e ipostatizzate. (Lévy, 1994, trad. it., p. 34)

Per Lévy la metafora più adatta per rappresentare il funzionamento dell'intelligenza collettiva non è quella del formicaio, della società totalitaria in cui gli individui sono asserviti al bene comune, ma quella di una «cosmopedia» pulsante e multidimensionale continuamente alimentata dalla collaborazione tra individui. L'enciclopedia universale frutto dell'intelligenza collettiva descritta da Lévy sembra esattamente richiamare, in maniera profetica, ciò che è avvenuto quasi dieci anni dopo nel progetto Wikipedia.

> I membri di una comunità pensante cercano, iscrivono, connettono, consultano, esplorano. Il loro sapere collettivo si materializza in una immensa immagine elettronica pluridimensionale, in continua metamorfosi che germoglia al ritmo delle invenzioni, delle scoperte, quasi vivente. La cosmopedia non solo mette a disposizione dell'intellettuale collettivo l'insieme delle conoscenze esistenti e pertinenti per lui in un dato momento, ma rappresenta anche un luogo fondamentale di discussione, negoziazione ed elaborazione collettiva. […] Colui o colei che si limita a consultare una questione di biochimica o di storia dell'arte sarà in grado di inserire nuovi enunciati riguardanti un certo setto-

---

che ne scaturisca una vita più alta, la quale, reagendo sugli elementi da cui risulta, li eleva a una forma superiore di esistenza e li trasforma». (Durkheim, 1912, trad. it., p. 510).

re dell'elettronica o dello svezzamento, di cui (lui o lei) è specialista. (Lévy, 1994, trad. it., p. 211)

Il sociologo dei media americano Henry Jenkins, in *Fans, Bloggers, and Gamers* (2006), richiama esplicitamente il concetto di intelligenza collettiva proposto da Lévy per analizzare la dimensione culturale e sociale delle comunità online di fan e appassionati dei media. Le numerose comunità composte da seguaci di serie televisive, di star della musica e del cinema, di sport, di videogiochi ecc. sono uno dei bacini privilegiati per lo studio delle nuove dinamiche sociali supportate ed influenzate dal mezzo Internet.

Online fan communities might well be some of the most fully realized versions of Levy's cosmopedia, expansive self-organizing groups focused around the collective production, debate, and circulation of meanings, interpretations, and fantasies in response to various artifacts of contemporary popular culture. (Jenkins, 2006[2], p. 137)

I fan sono invogliati a condividere le proprie conoscenze per ottenere rispetto e visibilità all'interno di una comunità costituita da persone che condividono i loro stessi interessi. Sanno di far piacere agli altri mettendo a disposizione le loro scoperte (un aneddoto, una fotografia, un link, ecc) in quanto essi stessi provocano piacere nel ricevere questi doni. Entra in gioco dunque un meccanismo di identificazione che è il vero collante delle comunità di giocatori e fan online.

Soap talk, Baym notes, allows people to "show off for one another" their various competencies while making individual expertise more broadly available. Fans are motivated by epistemaphilia - not simply a pleasure in knowing but a pleasure in exchanging knowledge. Baym argues that fans see the exchange of speculations and evaluations of soaps as a means of "comparing, refining, and negotiating understandings of their socioemotional environment." (Jenkins, 2006 [2], p. 139)

In Jenkins è molto evidente la centralità della cultura collaborativa come mezzo per attivare i meccanismi dell'intelligenza collettiva. Quando parla specificamente di Wikipedia in *Convergence Culture* (2006), la considera infatti come uno dei frutti del movimento FOSS (Free and Open Source) che ha avuto origine agli inizi degli anni Ottanta in ambito informatico e che poi ha preso sempre più piede nella società americana ed europea come modello di produzione e condivi-

sione della cultura. Sia la comunità FOSS che quella dei wikipediani sono organizzate come delle *adhocracies*, ovvero modelli di organizzazione non gerarchici, dinamici, organici, dove ognuno è spinto a contribuire secondo le proprie peculiari capacità.

> So far, adhocracy principles have been embraced by the open-source movement , where software engineers worldwide collaborate on projects for the common good. The Wikipedia project represents the application of these open-source principles to the production and management of knowledge. [...] Perhaps the most controversial aspect of the Wikipedia project has been the ways it shifts what counts as knowledge (from the kind of topics sanctioned by traditional encyclopedias to a much broader range of topics relevant to specialized interest groups and subcultures) and the ways it shifts what counts as expertise (from recognized academic authorities to something close to Lévy's concept of collective intelligence). (Jenkins, 2006 [2])

James Surowiecki intende al contrario per intelligenza collettiva qualcosa di più specifico, che non prevede alcun meccanismo di collaborazione o di contatto tra i componenti della collettività. Per il giornalista americano, autore di *The Wisdom of Crowds* (2004), l'intelligenza collettiva è quella caratteristica che si manifesta quando la media dei giudizi di tutti i membri di un gruppo si avvicina notevolmente al giudizio della persona più esperta, ovvero al giudizio migliore in quella circostanza. Il giudizio, la soluzione o la previsione collettiva, che è la media di giudizi, soluzioni o previsioni date singolarmente e senza alcun tipo di comunicazione orizzontale, manifesta una sorprendente "intelligenza", tanto da essere chiamata dall'autore "saggezza della folla". Gli esempi forniti da Surowiecki sono molteplici e tratti da diversi ambiti. Quello forse più noto, perché particolarmente intuitivo, è l'esperimento dello scienziato inglese Francis Galton condotto nel 1906. Trovandosi a una fiera contadina, Galton ha assistito ad una competizione dove al pubblico veniva richiesto di indovinare il peso di un bue. Ogni singola persona, contadini, commercianti, semplici spettatori, scriveva la sua previsione su un pezzetto di carta, che veniva inserito dagli organizzatori in una grande urna. Ovviamente il premio (il bue stesso) sarebbe andato a colui che avesse indovinato il peso esatto. Alla fine della gara, Galton, curioso di conoscere la distribuzione dei giudizi, si è fatto consegnare l'urna contenente i 787 foglietti.

Dopo un calcolo statistico, è risultato che la media di tutte le previsioni si avvicinava in maniera impressionante al peso esatto del bue (1,197 libbre era la previsione, 1,198 libbre il peso reale). Altri esempi sono nell'ambito della finanza e della tecnologia. A seguito del disastro dello Space Shuttle Challenger, avvenuto nel 1986, i mercati hanno reagito facendo precipitare di molti punti le azioni della Thiokol, l'azienda costruttrice dei razzi a propellente solido utilizzati dalla nave spaziale. Ancora non si sapeva nulla sulle vere cause del disastro, dunque la responsabilità per quanto accaduto non era stata attribuita con certezza tra le diverse aziende che avevano contribuito alla costruzione del Chakkenger. Tuttavia la maggior parte degli azionisti hanno scelto di vendere le azioni della Thiokol, e la loro valutazione si è rivelata poco dopo essere corretta. Il disastro era stato causato in effetti da un difetto di fabbricazione degli O-Ring dei razzi a propellente solido, motivo per cui la Thiokol ha dovuto affrontare un procedimento penale e ha perso del tutto credibilità e mercato. Un caso evidente di intelligenza collettiva, in campo tecnologico, è per Surowiecki il motore di ricerca Google. Google si basa su un algoritmo di Information Retrieval chiamato PageRank (Page et al., 1998), che ordina i risultati della ricerca in base sia all'attinenza del documento con la query di ricerca sia in base alla rilevanza della pagina all'interno della Rete. Questa rilevanza è calcolata per mezzo del numero di *backlink* che tale pagina riceve da altre pagine Web, dunque è il frutto dell'intelligenza collettiva della Rete, che opera in sottofondo per far emergere determinate pagine e nasconderne altre. La straordinaria precisione dei risultati di Google è per Surowiecki la prova che questa forma di intelligenza di massa funziona e che l'azienda di Brin e Page non ha fatto altro che comprendere quanto potesse essere utile sfruttarla per fini commerciali.

Ciò che distingue la prospettiva di Surowiecki da quella di altri studiosi è la negazione della componente interazionale e comunicativa dell'intelligenza collettiva. La saggezza della folla non è funzione della sua capacità di far entrare in contatto i propri componenti e di portarli alla collaborazione:

> The four conditions that characterize wise crowds: diversity of opinion (each person should have some private information, even if it's just an eccentric interpretation of the known facts), independence (people's opinions are not determined by the opinions

> of those around them), decentralization (people are able to specialize and draw on local knowledge), and aggregation (some mechanism exists for turning private judgments into a collective decision). If a group satisfies those conditions, its judgements is likely to be accurate. (Surowiecki, 2004, p. 10)

Non vi è un meccanismo specifico, una forma di *networking* o un metodo sociale per aumentare l'intelligenza collettiva, questa sembra invece essere qualcosa di radicato nella natura umana, come se la specie umana fosse programmata per agire come un unico cervello collettivo perfettamente efficiente. Da questo punto di vista, la posizione di Surowiecki ispira un certo grado di immanentismo dalle interessanti ricadute teoriche, che tuttavia non possono essere argomento di questa tesi:

> In cases like Francis Galton's experiment or the Challanger explosion, the crowd is holding a nearly complete picture of the world in its collective brain. [...] With most things, the average is mediocrity. With decision making, it's often excellence. You could say it's as if we've been programmed to be collectively smart. (Surowiecki, 2004, p. 11)

Al contrario, lo scienziato e divulgatore canadese Micheal Nielsen, nel suo recente saggio *Reinventing Discovery: The New Era of Networked Science* (2012), si focalizza sugli strumenti che le tecnologie dell'informazione, e soprattutto Internet, mettono oggi a disposizione per amplificare l'intelligenza collettiva. La scienza sta vivendo, secondo Nielsen, una rivoluzione radicale caratterizzata dal passaggio da una dimensione individualista della ricerca e della scoperta a una concezione più collaborativa e aperta al contributo delle masse. Mettere i propri dati a disposizione sia degli altri scienziati sia di figure professionali diverse (per esempio informatici e artisti) è una fonte enorme di arricchimento, perché favorisce utilizzi inaspettati dei dati stessi. Integrare formati e processi tra i gruppi di ricerca di uno stesso settore aiuta il reciproco scambio di informazioni e di esperienze.

In particolar modo, per la soluzione di problemi complessi, l'intelligenza collettiva all'interno di gruppi estesi e disomogenei è di sorprendente utilità. Un esempio portato da Nielsen è la celebre partita di scacchi "Kasparov vs. the

World" giocatasi su Internet nel 1999. L'evento fu organizzato da Microsoft per pubblicizzare la piattaforma di gioco online MSN Gaming Zone. Chiunque fosse interessato a partecipare alla sfida contro il campione del mondo di scacchi in carica, poteva iscriversi al portale dell'evento e di volta in volta esprimere con una votazione la propria preferenza per la mossa da effettuare. La compagine del World Team era composta da più di 50.000 persone di 75 paesi. Il World Team aveva a disposizione vari strumenti tecnologici per comunicare, per visualizzare l'andamento della partita, per scambiarsi analisi delle varianti in tempo reale, ecc. Fin dalla prima mossa, il forum per la discussione brulicava di pareri, scambi di pareri e di analisi. Una società informatica chiamata Smart Chess aveva messo a disposizione un software di visualizzazione dell'albero delle varianti che mostrava le mosse e le contromosse possibili a favore o contro le diverse linee di gioco. Questo, insieme all'azione di quattro coordinatori scelti da Microsoft (giovani scacchisti emergenti), facilitò molto l'organizzazione della squadra al proprio interno. La composizione del Word Team era tutt'altro che uniforme: vi erano sicuramente Maestri e giocatori professionisti, stuzzicati dall'originalità della sfida, ma la maggior parte erano dilettanti e semplici appassionati. Addirittura il 2,4% era composto da persone che non conoscevano nemmeno le regole del gioco, come dimostrato dai risultati delle votazioni (Nielsen, 2012, p. 20). Nonostante questo, la partita si mantenne equilibrata fino alla fine. Solo alla cinquantunesima Kasparov riuscì ad avvantaggiarsi, per andare a vincere, dopo un finale di partita sofferto, alla sessantaduesima. Kasparov definì l'incontro come una delle partite di scacchi più complesse ed esaltanti della sua carriera, ne nacquero soluzioni che ancora oggi sono considerate rivoluzionarie[25].

Qual è stato il motivo di tale successo? Il fatto che il World Team fosse composto da così tante persone è stata la carta vincente? Secondo Nielsen, no. A prova del fatto che non si è trattato solo di una questione numerica, lo studioso cita un'analoga impresa tentata tre anni prima contro il campione russo Anatoly Karpov :

[25] Le parole di Kasparov furono: «It is the greatest game in the history of chess. The sheer number of ideas, the complexity, and the contribution it has made to chess make it the most important game ever played.» (Harding et al., 2002, p. 98)

> "Karpov against the World" used a different online system to decide moves, with no game forum or official game advisors, and giving World Team members just ten minutes to vote on their preferred move. Without the means to coordinate their actions, the World Team played poorly, and Karpov crushed them in just 32 moves. (Nielsen, 2012, p. 18)

La chiave di volta non è dunque il numero, ma il metodo. Sono gli strumenti per la collaborazione e l'interazione tra i membri del gruppo a fornire un vantaggio competitivo alla squadra. Quel che Nielsen vuole dimostrare è che l'intelligenza collettiva può essere favorita dalle tecnologie dell'informazione e della comunicazione, e che nell'ambito della scienza questa è un'occasione da non perdere. L'analogia col progetto Wikipedia è abbastanza evidente. La tecnologia del Wiki (la piattaforma che permette a più utenti di modificare contemporaneamente in tempo reale una pagina Web) assieme a strumenti come le sezioni "Talk" e "View history", hanno favorito la riuscita del modello collaborativo di Wikipedia. Questi strumenti hanno fornito la base per l'interazione proficua dei componenti della comunità proprio come i forum e gli alberi decisionali hanno aiutato il World Team a tenere testa a Garry Kasparov. Per Lévy un aspetto fondamentale dell'intelligenza collettiva è che essa riesce a «mobilitare in maniera effettiva le competenze» (Lévy, 1994, trad. it., p. 35). Questo concetto è approfondito da Michael Nielsen, che spiega come le logiche di gruppo stimolino un'attivazione delle micro-competenze latenti («harnessing latent microexpertise», Nielsen, 2012, p. 23) dei singoli individui. Durante una partita di scacchi, per esempio, le diverse fasi di gioco (apertura, mediogioco, finale) vedono susseguirsi numerosi elementi tattici e strategici tipici della particolare impostazione che i giocatori danno alla partita. Un tipo di apertura può dar seguito ad un insieme di varianti e di mosse tattiche che conducono a un mediogioco più o meno "aperto", che può a sua volta svilupparsi in un certo genere di finale. Data la complessità del gioco e il numero di possibilità presenti, è praticamente impossibile per un giocatore, anche un Grande Maestro, essere specializzato in tutti i tipi di partita. Nel circuito agonistico infatti esistono perlopiù "specialisti" in determinati tipi di gioco (l'apertura Est Indiana, la difesa Pirc, il gambetto di Donna, ecc) che cercano di mettere in difficoltà gli altri giocatori esperti condu-

cendoli nel proprio terreno di gioco preferito. È evidente che giocatori del livello di Kasparov mantengono un'elevatissima qualità di gioco in molte diverse varianti, ma anche loro hanno competenze maggiori in alcune e minori in altre. Il grande vantaggio del modello collaborativo di gioco sperimentato durante Kasparov vs the World è che riesce a sfruttare nelle diverse fasi della partita le diverse competenze dei singoli. Nessuno dei 50.000 giocatori del World Team aveva in tutte le fasi della partita competenze del livello di Kasparov, ma nelle specifiche fasi di gioco era probabile che almeno un giocatore avesse competenze del suo livello:

> The key to the World Team's play was to ensure that all this ordinarily latent microexpertise was uncovered and acted upon in response to the contingencies of the game. So although it was a lucky chance that Krush [una giovane ma geniale giocatrice di scacchi del World Team] in particular was the person whose microexpertise was decisive at move 10, given the number of experienced chess player involved, it was highly likely that latent microexpertise from those players would come to light at critical points during the game, and so help the World Team match Kasparov. (Nielsen, 2012, p. 25)

Nella comunità di Wikipedia accade un fenomeno simile. Un utente comincia a scrivere una voce in cui si sente abbastanza esperto, mettendo a disposizione le proprie competenze. Dopo un certo periodo di tempo, esiste un'alta probabilità che altri utenti, che hanno microcompetenze specifiche in quell'argomento o in sottoinsiemi di quell'argomento, comincino ad aggiungere pezzi, a correggere informazioni, ad arricchire i riferimenti e la bibliografia. Ognuno offre il suo piccolo contributo in base al suo piccolo bagaglio di conoscenza: in una comunità di 35 milioni di utenti registrati è facile che si riesca a trovare almeno un contributore abbastanza esperto per ogni singolo argomento. Le diverse passioni e bizzarrie di ognuno spaziano nei più svariati campi (i fumetti giapponesi, le monete antiche, le serie televisive, la paleontologia, le mappe, l'ortografia, ecc.) e trovano facilmente un riscontro in quell'amalgama di tematiche che è Wikipedia, con l'effetto ulteriore di integrare o ispirare le attività degli altri (Lih, 2009).

Per cultura collaborativa si intende un insieme di assunzioni, valori, significati e azioni che pertengono al lavorare insieme in una comunità. Le comunità ba-

sate su un principio di collaborazione hanno barriere d'ingresso basse e una struttura debolmente gerarchica: favoriscono invece i rapporti orizzontali tra i singoli membri, gli scambi alla pari e la condivisione di significati, esperienze e competenze.

> Collaboration is the process of shared creation: two or more individuals with complementary skills interacting to create a shared understanding that none had previously possessed or could have come to on their own. Collaboration creates a shared meaning about a process, a product, or an event. (Schrage, 1990, p. 40.)

Non stupisce che questo concetto sia stato chiamato in causa spesso per definire il modello della comunità informatica Free and Open Source. FOSS è un movimento sociale, partito nell'ambito della programmazione, che rifiuta la segretezza e il controllo centralizzato del lavoro creativo in favore della trasparenza e della condivisione senza restrizioni dell'informazione. Le radici dell'*open source* risalgono alle pratiche informatiche *in auge* negli anni Sessanta nel mondo accademico e nelle comunità di programmatori sia in America che in Europa. Gli sviluppatori spesso condividevano tra loro il proprio codice in modo informale e lo modificavano collettivamente per risolvere i problemi tecnici. La pratica di condividere il codice sorgente è stata molto efficace e coerente all'interno della comunità di sviluppo del sistema operativo UNIX, tanto da essere una delle ragioni iniziali del suo successo. Il passaggio dalla condivisione informale di codice all'esplicita concettualizzazione del modello *open source* si è avuto agli inizi degli anni Ottanta grazie all'opera di Richard Stallman. Stallman era un carismatico programmatore del Massachusetts Institute of Technology che decise di intraprendere una battaglia a viso aperto contro le grandi *software house* che volevano mantenere il software chiuso e proprietario: nel 1984 rassegnò le dimissioni dal MIT per fondare il progetto GNU con l'obiettivo di sviluppare un sistema operativo UNIX-like completamente libero e *open source*. È a Stallman che si deve la fondazione del movimento FOSS e della Free Software Foundation[26], di cui è ancora presidente.

[26] URL: http://www.fsf.org/

All'interno delle comunità di sviluppo *open source* si è diffusa quella che Manuel Castells, in *The Internet Galaxy* (2001), chiama «cultura hacker». Per Castells alle origini di Internet ci sono quattro fondamentali agenti culturali che ne hanno plasmato la forma e i contenuti: la cultura tecno-meritocratica delle università, la cultura hacker, quella comunitario virtuale e quella imprenditoriale[27]. Gli hacker mettono in Rete il loro contributo allo sviluppo del software confidando nel principio di reciprocità. La loro è una cultura che riconosce grande valore al merito, in quanto è finalizzata alla creazione di progetti software innovativi, ma anche alla collaborazione come metodologia migliore per la soluzione dei problemi[28]. Riuscire a scrivere del codice utile o aiutare gli altri a farlo è il mezzo con cui gli hacker ottengono rispetto e gratitudine in una comunità di pari, e per questo le logiche del profitto economico passano in secondo piano.

> The gift culture in the hacker world is specific vis-à-vis other gift cultures. Prestige, reputation and social esteem are linked to the relevance of the gift to the community. So, it is not only the expected return for generosity, but the immediate gratification of displaying to everybody the hacker's ingenuity. In addition, there is also gratification involved in the object of the gift. It not only has exchange value, but also use value. The recognition comes not only from giving, but from producing a valuable object (innovative software). (Castells, 2001, p. 47.)

La tematica del dono è riconosciuta come centrale nella cultura collaborativa anche da Raffaele Meo e Mariella Berra (2006). Per gli autori i modelli etico-comportamentali alla base delle comunità hacker sono vicini a quelli delle società arcaiche pre-capitaliste in cui prevalevano logiche di scambio e di reciprocità piuttosto che di mercato. Il dono delle proprie conoscenze o del frutto del proprio

---

[27] Anche Howard Rheingold in *Smart Mobs: The Next Social Revolution*, riconosce alle comunità hacker un ruolo di primo piano nella nascita e nello sviluppo iniziale di Internet: «The Internet was deliberately designed by hackers to be an innovation commons, a laboratory for collaboratively creating better technologies. They knew that some community of hackers in the future would know more about networks than the original creators, so the designers of the Internet took care to avoid technical obstacles to future innovation» (Rheingold, 2002, p. 48).

[28] Come scrive Eric Raymond nel saggio *The Cathedral and the Bazaar*, «The open-source peer review is the only scalable method for achieving high reliability and quality» (Raymond, 1999, p. 170).

lavoro tra gli hacker non è motivato da un calcolo utilitaristico o contrattuale, da un *do ut des*, ma dall'etica sociale dove il dono è riconosciuto come dovere e senso della comunità stessa. In quest'ottica non è tanto la qualità o la quantità del dono a contare, ma l'atto stesso del donare, che diventa un riconoscimento dell'appartenenza al gruppo e della coesione tra i suoi membri.

> Nel dono non sono le leggi del mercato e della concorrenza a regolare i rapporti fra le persone, né le regole formali imposte da una gerarchia come nella redistribuzione, ma donatori e donatari sono vincolati da una obbligazione morale allo scambio reciproco inscritto in una prescrizione sociale. Mentre nel commercio lo scambio è immediato e viene regolato attraverso il contratto, il dono disegna un rapporto che continua in un ciclo infinito. (Berra et al., 2006, p. 167)

Nel celebre articolo del 2002 *Coase's Penguin. Linux and The Nature of the Firm*, Yochai Benkler indica Wikipedia come uno degli esiti più riusciti della cultura collaborativa delle comunità sul Web. Egli definisce come "produzione tra pari basata sul bene comune" il fenomeno in base al quale diverse persone lavorano insieme per lo stesso obiettivo, che sia scrivere codice o più in generale creare conoscenza libera di essere copiata, distribuita, utilizzata e modificata da altri. Secondo Benkler, se la retribuzione economica e la costituzione di compagnie societarie sono comunemente riconosciute come le spinte propulsive per l'innovazione e il progresso umano, ci dev'essere qualcos'altro a ispirare i volontari che lavorano per Linux, Wikipedia e gli altri progetti "liberi" di grande successo nell'età dell'informazione. Egli afferma che queste motivazioni scaturiscono da due fattori estranei al denaro: l'appagamento sociale e psicologico che si ottiene dal relazionarsi con gli altri e la gratificazione edonistica propria del compito che ci si prefigge (considerato eticamente giusto e "nobile").

> The incentive problem as an objection to the general sustainability of peer production is in large part resolved by the existence of a series of mechanisms for indirect appropriation of the benefits of participation. At the broadest level, there is the pleasure of creation. Whether you refer to this pleasure dispassionately as "hedonic gain" or romantically as "an urge to create," the mechanism is simple. People are creative beings. They will play at creation if given an opportunity, and the network and free access to information resources provide this opportunity. [...] It becomes relatively straightfor-

> ward to see that there will be conditions under which a project that can organize itself to offer social-psychological rewards removed from monetary rewards will attract certain people, or at least certain chunks of people's days, that monetary rewards would not. (Benkler, 2002, p. 378 e 424)

Le influenze della cultura *open source* sulla nascita di Wikipedia sono riconosciute dalla maggior parte degli studiosi del fenomeno Wikipedia (Ayers et al., 2008, Lih, 2009; Reagle Jr., 2010, ecc.). D'altronde lo stesso fondatore Jimmy Wales ne ha più volte riconosciuto il debito. Wales ha affermato[29] che un articolo divulgativo di Eric Raymond, che collegava la produzione FOSS a quella di un bazar vibrante e decentrato, gli ha aperto gli occhi sulle opportunità della collaborazione di massa. Nell'ottobre del 2001, quando Wikipedia non aveva ancora un anno di vita, Wales raccolse in un elenco i principi che riteneva fossero responsabili del successo dell'*open source* e sperava sarebbero stati determinanti per la riuscita di Wikipedia. Gli Statements of Principles dichiaravano tra le altre cose che il successo di Wikipedia poteva essere solo funzione della sua comunità e della natura aperta e collaborativa della sua produzione. (Reagle, 2010, p. 78).

Oltre alla cultura del dono, Wikipedia eredita dalla componente hacker anche i principi di trasparenza, non-discriminazione ed "eventualismo". La trasparenza in Wikipedia è soprattutto di processo: le sezioni "Talk" e "View history" di ogni pagina dell'enciclopedia mostrano gli sviluppi tortuosi di ogni singola voce e testimoniano il meccanismo di costruzione del consenso. Gli scambi di opinioni e i dibattiti tra i wikipediani sono pubblici e ognuno può contribuirvi; così come sono disponibili tutte le passate versioni di una voce e le motivazioni per cui sono state modificate, cancellate o ripristinate. La trasparenza consente ai partecipanti ai progetti *open* di capire le ragioni dietro a ogni singola decisione, contribuendo a creare fiducia nel processo di Wikipedia. Fornisce inoltre ai *newbies* (gli ultimi arrivati) uno strumento per comprendere la cultura e i protocolli informali della comunità, riducendo i comportamenti inappropriati. La tecnologia del Wiki è stata progettata proprio in vista della trasparenza e della totale documentazione dei processi creativi o decisionali. Per non-discriminazione si intende il principio

[29] URL: http://www.newyorker.com/archive/2006/07/31/060731fa_fact

secondo cui ogni opinione in Wikipedia è ascoltata, anche quando non condivisa, indipendentemente dalle caratteristiche della persona che la esprime (provenienza geografica, razza, genere sessuale, orientamento politico o religioso, ecc.). Le persone e le loro azioni devono essere giudicate in base al merito, e il merito è stabilito dal livello qualitativo e quantitativo di contribuzione all'enciclopedia. L'ultimo principio, quello di "eventualismo" (*eventualism*[30]) è stato ravvisato da Andrew Lih nel suo *The Wikipedia Revolution* (2009). Esso deriva direttamente dalla logica *open source* e indica che il contributo iniziale a un progetto comunitario non dev'essere per forza perfetto o completo, ma può anche semplicemente essere "abbozzato". Questa bozza (*stub*[31] in termini wikipediani) sarà poi completata, migliorata, arricchita e corretta nel tempo dal resto della comunità, che vi si getterà sopra con un "effetto piranha":

> An article may not be great now, but even without a deadline, it will eventually be made better in the future by someone else. It was a sign of faith in the piranha effect taking hold, eventually. […] Eventualism has become an accepted norm in the community, because by default since the beginning of the project, starting from nothing, articles have overwhelmingly benefited from multiple eyeballs (and edit). Entries have generally increased in quality over time, giving more and more faith to the theory that articles by and large attract more content. (Lih, 2009, p. 141)

1.3 Criticità in Wikipedia

Dalla sua nascita ad oggi numerose sono state le critiche al progetto Wikipedia. Sebbene non vi sia ancora uno studio sistematico su questo argomento, diversi contributi al dibattito si possono trovare in forma sparsa su giornali o riviste, sia cartacei che digitali, su blog e paper scientifici[32]. È interessante rilevare che la maggior parte di questi contributi sono apparsi nel periodo dal 2004 al

[30] La traduzione "eventualismo" rende l'idea, ma è un po' forzata, in quanto in inglese l'aggettivo "eventual" significa "finale". Potrebbe essere reso parimenti con "finalismo" o "conclusivismo".

[31] URL: http://en.wikipedia.org/wiki/Wikipedia:Stub

[32] Un tentativo di raccogliere le "preoccupazioni" dell'opinione pubblica americana ed europea sul fenomeno Wikipedia è stato effettuato da Reagle Jr, nell'ultimo capitolo di *Good Faith Collaboration*, intitolato significativamente *Encyclopedic Anxiety*. (Reagle Jr, 2009, pp. 137-168).

2007, gli anni in cui il fenomeno Wikipedia emergeva all'attenzione del mondo accademico e dell'opinione pubblica senza avere ancora raggiunto un grado sufficiente di maturità. Si tratta prevalentemente di critiche che riguardano il contenuto e la forma degli articoli, ma anche più in generale il processo di creazione, validazione e revisione dei contenuti, nonché la comunità stessa dei wikipediani. Per procedere in maniera ordinata, queste argomentazioni saranno passate in rassegna in base al difetto principale che esse individuano in Wikipedia o nella sua community:

1) inaccuratezza e inaffidabilità;
2) copertura diseguale dei diversi ambiti del sapere;
3) parzialità ed esposizione a pregiudizi e interessi;
4) volatilità.

1.3.1 Inaccuratezza e inaffidabilità

Essendo il primo obiettivo di una enciclopedia quello di essere una risorsa di consultazione accurata e affidabile, questo genere di critica punta al cuore di Wikipedia come opera di consultazione. La critica di Robert McHenry (*editor-in-chief* dell'Enciclopedia Britannica dal 1992 al 1997), per esempio, mette in dubbio la razionalità e l'efficacia del processo di creazione "collaborativa" della conoscenza che impronta Wikipedia. In un articolo per il settimanale online Ideas In Action[33] intitolato *The Faith-Based Encyclopedia*, McHernry parte con l'individuazione di errori e inconsistenze nella pagina di Wikipedia riguardante Alexander Hamilton:

> To see what Wikipedia is like I chose a single article, the biography of Alexander Hamilton. I chose that topic because I happen to know that there is a problem with his birth date, and how a reference work deals with that problem tells me something about its standards. The problem is this: While the day and month of Hamilton's birth are known, there is some uncertainty as to the year, whether it be 1755 or 1757. Hamilton himself used, and most contemporary biographers prefer, the latter year; a reference work ought at least to note the issue. The Wikipedia article on Hamilton (as of November 4, 2004) uses the 1755 date without comment. Unfortunately, a couple of refer-

[33] URL: http://www.ideasinactiontv.com/

> ences within the body of the article that mention his age in certain years are clearly derived from a source that used the 1757 date, creating an internal inconsistency that the reader has no means to resolve. Two different years are cited for the end of his service as secretary of the Treasury; without resorting to another reference work, you can guess that at least one of them is wrong. The article is rife with typographic errors, styling errors, and errors of grammar and diction. No doubt there are other factual errors as well, but I hardly needed to fact-check the piece to form my opinion. The writing is often awkward, and many sentences that are apparently meant to summarize some aspect of Hamilton's life or work betray the writer's lack of understanding of the subject matter. (McHernry, 2004)

Dopo aver evidenziato questi errori nella versione allora corrente (novembre 2004) dell'articolo, McHenry scorre le versioni precedenti scoprendo che molte di esse risultavano più corrette e scritte meglio. Per McHenry questa è una prova del fatto che aprire la conoscenza alla massa degli utenti e mantenere il contenuto aperto a ulteriori modiche nel tempo non è affatto garanzia che tale contenuto migliori, in quanto non è detto che l'opinione degli utenti più esperti e competenti in materia prevalga sulla massa degli «uninformed and semiliterate muddlers» (pasticcioni disinformati e semianalfabeti). Alla base di Wikipedia ci sarebbe dunque un principio ingenuo e infondato, fideisticamente assunto per vero da una comunità forse troppo infarcita di filosofia 2.0:

> Some unspecified quasi-Darwinian process will assure that those writings and editings by contributors of greatest expertise will survive; articles will eventually reach a steady state that corresponds to the highest degree of accuracy. Does someone actually believe this? […] Take the statements of faith in the efficacy of collaborative editing, replace the shibboleth "community" with the banal "committee," and the surprise dissolves before your eyes. Or, if you are of a statistical turn of mind, think a little about regression to the mean and the shape of the normal distribution curve. (McHernry, 2004)

Lo storico Roy Rosenzweig, invece, mette in evidenza la mancanza di profondità e di coerenza interna di Wikipedia, con particolare riferimento alle voci riguardanti argomenti storici. In molti casi lo studioso rileva prosa inelegante e analisi debole, una struttura confusa che rasenta l'incoerenza e perfino il plagio di altre risorse presenti online. Per quanto riconosca i meriti di Wikipedia come

strumento di consultazione online (meglio per esempio di Encarta, il corrispondente prodotto commerciale di Microsoft), Rosenzweig afferma che Wikipedia non regge il confronto con i compendi scritti da storici professionisti, quali la American National Biography Online[34]:

> Good historical writing requires not just factual accuracy but also a command of the scholarly literature, persuasive analysis and interpretations, and clear and engaging prose. By those measures, American National Biography Online easily outdistances Wikipedia. (Rosenzweig, 2006, p. 64)

Il difetto principale di Wikipedia è la mancanza di una contestualizzazione delle singole voci in un'analisi di più ampio respiro, ovvero la carenza di un'interpretazione in grado di andare al di là del semplice collage di informazioni trovate su fonti sparse. I contributori dell'enciclopedia non sono capaci, come lo storico professionista, di sintetizzare le informazioni in un quadro organico che sia nello stesso tempo istruttivo e piacevole da leggere. La mancanza di un punto di vista ben definito (peraltro esplicitamente consigliata da Wikipedia nel principio Neutral Point of View) è vista da Rosenzweig come un difetto, perché rischia di ridurre gli articoli a un'inutile pappardella («waffling») che espone pedissequamente le diverse posizioni su un argomento come avessero tutte la stessa rilevanza.

> Wikipedia's profile of the Confederate guerrilla fighter William Clarke Quantrill arguably does a better job of detailing the controversies about his actions than American National Biography Online. Even so, it provides a typical waffling conclusion that contrasts sharply with the firm judgments in the best of the American National Biography Online essays: "Some historians," they write, "remember him as an opportunistic, bloodthirsty outlaw, while others continue to view him as a daring soldier and local folk hero." This waffling—encouraged by the npov policy—means that it is hard to discern any overall interpretive stance in Wikipedia history. (Rosenzweig, 2006, p. 65)

La critica di un *insider* come Larry Sanger, cofondatore di Wikipedia insieme a Jimmy Wales[35] è essenziale per comprendere le diverse istanze alla base

[34] URL: http://www.anb.org/orderinginfo.html

[35] Sanger ha abbandonato il progetto nel 2002 per divergenze sulla politica editoriale.

dell'enciclopedia. Fin dall'inizio Wales e Sanger manifestarono un diverso modo di concepire il progetto, perché laddove Wales insisteva sull'apertura dei contenti e sulla parità tra i membri della community, Sanger proponeva di favorire i contributi degli esperti, sia nella produzione di contenuti che nella soluzione delle divergenze. Il principale difetto di Wikipedia sottolineato da Sanger è la scarsa credibilità agli occhi dei lettori più esigenti, dovuta all'assenza di un controllo editoriale da parte di personale esperto:

> When it comes to relatively specialized topics (outside of the interests of most of the contributors), the project's credibility is very uneven. If the project was lucky enough to have a writer or two well-informed about some specialized subject, and if their work was not degraded in quality by the majority of people, whose knowledge of the subject is based on paragraphs in books and mere mentions in college classes, then there might be a good, credible article on that specialized subject. Otherwise, there will be no article at all, a very amateurish-sounding article, or an article that looks like it might once have been pretty good, but which has been hacked to bits by hoi polloi. (Sanger, 2004)

Sanger spiega che all'interno della comunità dei wikipediani c'è un sentimento diffuso ed esplicito di "anti-elitarismo", a causa del quale il contributo degli esperti non solo non è tenuto in maggiore considerazione rispetto a quello degli utenti comuni, ma è anzi snobbato o addirittura osteggiato.

> Namely, as a community, Wikipedia lacks the habit or tradition of respect for expertise. As a community, far from being elitist (which would, in this context, mean excluding the unwashed masses), it is anti-elitist (which, in this context, means that expertise is not accorded any special respect, and snubs and disrespect of expertise is tolerated). (Sanger, 2004)

Senza l'apporto di docenti, studiosi, ricercatori e in generale di esperti nelle specifiche materie, Wikipedia non può superare i suoi limiti: per questo Sanger auspica un cambiamento nelle policy del progetto che consenta di attirare scrittori qualificati disincentivando contemporaneamente perditempo e disturbatori («trolls»). Un meccanismo potrebbe essere quello di attribuire riconoscimenti e premi agli autori che si sono dimostrati più competenti ed hanno contribuito maggiormente al miglioramento dell'enciclopedia.

Di fatto Sanger, dopo aver abbandonato il gruppo di Wikipedia, ha realizzato tale disegno fondando Citizendium[36], un'enciclopedia online dove maggior peso è attribuito al contributo degli esperti e vi è un maggiore controllo editoriale sui contenuti. Citizendium è stato concepito inizialmente come un *fork* del corpus di Wikipedia, con l'obiettivo di impiantare nuovi contenuti su quelli già presenti. Proposito presto abbandonato per problemi legati alla licenza. "Forkare" da Wikipedia significava infatti ereditarne la licenza GFDL e dunque consentire a chiunque di reimportare i contenuti migliorati da Citizendium a Wikipedia, annullando di fatto la differenza qualitativa tra le due enciclopedie.

1.3.2 Copertura diseguale dei diversi ambiti del sapere

Nel suo contributo *Can History be Open Source? Wikipedia and the Future of the Past*, Rosenzweig sottolinea anche lo squilibrio presente in Wikiepdia, a livello sia quantitativo sia qualitativo, tra le diverse voci e i diversi campi del sapere. Per Rosemberg questo squilibrio rispecchia gli interessi della comunità dei wikipediani e tende a favorire gli argomenti più popolari nella cultura di massa rispetto a quelli accademici e scientifici.

> Participation in Wikipedia entries generally maps popular, rather than academic, interests in history. U.S. cultural history, recently one of the liveliest areas of professional history writing, is what Wikipedia calls a "stub" consisting of one banal sentence ("The cultural history of the United States is a broad topic, covering or having influence in many of the world's cultural aspects."). By contrast, Wikipedia offers a detailed 3,100-word article titled "Postage Stamps and Postal History of the United States," a topic with a devoted popular following that attracts little scholarly interest. (Rosenzweig, 2006, p. 60)

Sono le tematiche adolescenziali o vicine all'ambiente *geek* (informatica, fantascienza, hobbies vari, sessualità, televisione) ad avere una posizione di vantaggio rispetto a quelle della cultura tradizionale, come la storia, la letteratura o

[36] URL: http://en.citizendium.org/

l'arte. Citando il contenuto autocritico di una pagina di Wikipedia[37], Rosenzweig riporta:

> Geek priorities have shaped the encyclopedia: "There are many long and well-written articles on obscure characters in science fiction/fantasy and very specialised issues in computer science, physics and math; there are stubs, or bot [machine-generated] articles, or nothing, for vast areas of art, history, literature, film, geography." One regular contributor to Wikipedia's history articles observed (somewhat tongue in cheek): "Wikipedia kicks Britannica's ass when it comes to online mmp [massively multiplayer] games, trading card games, Tolkieana and Star Wars factoids!" (Rosenzweig, 2006, p. 61)

Un problema affine, evidenziato da Rosenzweig, è quello del "localismo", ovvero la tendenza a privilegiare temi di carattere locale, in cui porzioni minime di utenti hanno tuttavia un grande coinvolgimento, rispetto a tematiche di carattere generale, che coinvolgono un numero elevato di persone in maniera indiretta e superficiale. Perciò una voce che riguarda una cittadina dell'Oklahoma può avere lo stesso risalto dell'articolo su New York, se nella community c'è un gruppo di individui fortemente determinati ad approfondire ed arricchire i contenuti di quella voce per motivi affettivi o di appartenenza.

> "Articles tend to be whatever-centric," they [i wikipediani] acknowledge in one of their many self-critical commentaries. "People point out whatever is exceptional about their home province, tiny town or bizarre hobby, without noting frankly that their home province is completely unremarkable, their tiny town is not really all that special or that their bizarre hobby is, in fact, bizarre." That localism can sometimes cause conflicts on nonlocal entries, as in the Olmsted profile, where a Wikipedian from Louisville complains on the "Discussion" page that the biography overestimates Olmsted's work in Buffalo and ignores his work in—surprise!—Louisville. (Rosenzweig, 2006, p. 65)

Anche il filosofo britannico Martin Cohen evidenzia gli squilibri interni al corpus di Wikipedia tra le voci della cultura tradizionale e quelle del mondo giovanilistico. Facendo riferimento al paper del 2007 di Anselm Spoerri *What is Po-*

[37] URL: http://en.wikipedia.org/wiki/Wikipedia:Why_Wikipedia_is_not_so_great

pular in Wikipedia and Why?, Cohen deduce che l'intrattenimento è sicuramente il campo maggiormente trattato dagli autori di Wikipedia:

> Of the two and a half million articles in English, nearly half are in the "entertainment category", with science and the arts a miserly 6 per cent and 2 per cent respectively. (Cohen, 2008)

Il problema tuttavia per Cohen è di portata più vasta. Wikipedia, con i suoi articoli e le sue fonti di basso livello, sta influenzando negativamente generazioni di internauti, perpetuando luoghi comuni e pregiudizi dannosi per l'intera società. Quella contenuta in Wikipedia è la conoscenza media dei giovani maschi occidentali fanatici del computer: dare a questo tipo di sapere una dignità enciclopedica significa non consentire all'utente di Internet di andare al di là del superficiale e dell'erroneo, limitando di fatto la sua libertà di pensiero.

> Wikipedia's version of reality has already become a monopoly. And all the prejudices and ignorance of its creators are imposed too. To control the reference sources that people use is to control the way people comprehend the world. Wikipedia may have a benign, even trivial face, but underneath may lie a more sinister and subtle threat to freedom of thought. (Cohen, 2008)

Wikipedia mostra anche un certo grado di "autoreferenzialità digitale" dei contenuti. Le fonti citate si riferiscono raramente a testi cartacei non reperibili gratuitamente su Internet, in quanto più difficili da verificare per la comunità dei wikipediani. Tale difetto è conseguenza dello statuto stesso di Wikipedia, che da grande importanza alla controllabilità della fonte. *No Original Researches* significa che i contenuti più apprezzati sono quelli che si riferiscono a una fonte esterna, la quale per costituire una "prova" deve essere verificabile dagli altri utenti. Ciò porta inevitabilmente a privilegiare le fonti digitali in Rete, che sono le uniche a cui si può avere un accesso universale ed immediato. Se per esempio domani qualcuno avesse intenzione di aggiungere un paragrafo su Ramses III citando un testo conservato unicamente nella biblioteca di Alessandria, probabilmente il suo paragrafo troverebbe l'opposizione della comunità, perché, anche se vero, comunque sarebbe impossibile da verificare. Sicuramente si tratta di una critica

importante, perché non vi si scorge una soluzione compatibile con le policy editoriali di Wikipedia. La si ritrova in Cohen,

> On Wikipedia, knowledge is tracked instantly via Google searches, online newspapers and other internet encyclopaedias, not so much by consulting primary sources as "tertiary sources" - other internet sites. (Cohen, 2008)

e anche nel paper di Peter Denning, Jim Horning e altri intitolato *Wikipedia Risks*:

> Many articles do not cite independent sources. Few articles contain citations to works not digitized and stored in the open Internet. (Denning et al., 2005)

1.3.3 Parzialità ed esposizione a pregiudizi e interessi

Le critiche di parzialità a Wikipedia lamentano diversi tipi di pregiudizi interni alla comunità: americanismo, anti-conservatorismo, mediacentrismo, maschilismo.

Le posizioni filodemocratiche ed antirepubblicane di Wikipedia sono state sottolineate dal giornalista Rowan Scarborough in articolo per la rivista conservatrice americana *Human Events* dal titolo *Wikipedia Whacks the Right* (2010). Scarborough prende come esempi molti casi in cui le pagine dei candidati conservatori sono state ripetutamente editate dai wikipediani per aggiungere particolari negativi riguardanti la loro vita privata e la loro attività politica. Rispetto agli esponenti democratici, le posizioni dei conservatori sono volutamente estremizzate per accendere gli animi dei lettori e provocare sdegno. Ecco un esempio riguardante il diverso trattamento riservato da Wikipedia ai candidati al senato per il Delaware Coons (democratico) e O'Donnell (repubblicano):

> There is no balance in Wikipedia's treatment of Coons vs. O'Donnell. While Coons has a near-pristine career, according to Wikipedia, O'Donnell's page contains one controversy after another. O'Donnell's entry appears to be about ten times longer. It is filled with various controversies in her career, but little talk of achievements. It extensively lists her conservative positions on torch-hot issues such as abortion and English-only. The goal is clear: anger the Left and moderates. (Scarborough, 2010)

Per Scarborough gli autori delle voci dell'enciclopedia non soltanto sono di inclinazioni chiaramente *liberal*, ma il più delle volte attingono a fonti giornalistiche, come il New York Times e il Washington Post, che mostrano più severità verso i politici conservatori che verso i democratici:

> Because Wikipedia is driven by liberal-leaning contributors, entries for conservatives often come from the mainstream media—New York Times, Washington Post, major networks, et al. Since these news organizations criticize, investigate and scrutinize conservative Republicans much more often than they do Democrats, the pool of negative, thought not always accurate data is skewed. (Scarborough, 2010)

Proprio per contrastare e cercare di correggere nell'opinione pubblica le presunte tendenze democratiche propagandate da Wikipedia, nel 2006 è nato il progetto Conservapedia[38], un'enciclopedia online concorrente dalle spiccate posizioni conservatrici e cristiane.

Tim Anderson, *lecturer* alla School of Political Economy dell'Università di Sidney, denuncia la centralità in Wikipedia di fonti provenienti dai media istituzionali di lingua inglese (*corporate media*), come Time, CNN, Fox, BBC, ecc. Tali fonti vengono considerate dalla community più affidabili e neutrali delle altre, quando invece rispecchiano il punto di vista degli *opinion makers* dei paesi anglosassoni, se non addirittura interessi politici ed economici:

> US corporate media sources (Time, CNN, Fox, and so on) are privileged as reliable and "neutral" sources in Wiki entries, despite the fact that many of these bodies are intimately involved in many of the most contentious public debates, such as privatisation, intervention and war. […] The BBC was OK but ZNet and Venezuela Analysis were both unusable "biased" sources, unlike Time magazine. No "original research" was allowed but rather reportage based on administrator-determined "reliable" sources. (Anderson, 2008)

Il giornalista del New York Times Noam Cohen solleva la questione della disparità di genere: i contributori maschi nella community di Wikipedia sono in una percentuale molto superiore alle donne, che arrivano appena al 15% (Cohen, 2011). Questa disparità è rispecchiata a livello contenutistico dall'attenzione ri-

[38] URL: http://www.conservapedia.com/Main_Page

volta agli argomenti comunemente "maschili" (nell'esempio del testo le serie televisive *I Simpson*, *I Soprano* e il videogame *Grand Theft Auto*) rispetto a quelli che suscitano maggiore interesse per le donne (nell'esempio *Sex and The City* o i "braccialetti dell'amicizia"). L'articolo di Cohen ha dato il via a un certo dibattito attorno alla questione, come si può leggere nell'apposita "Room for Debate" ospitata dal sito del New York Times: *Where Are the Women in Wikipedia?*[39]

Le motivazioni del "maschilismo" di Wikipedia sono state per lo più individuate nel clima di conflitto, spesso accompagnato da toni accesi e aggressivi, che è alla base del processo di creazione collettiva delle singole voci. Essendo le donne meno a loro agio rispetto agli uomini nel tentare di far prevalere la loro voce, spesso rinuncerebbero a contribuire ad argomenti su cui avrebbero anche una preparazione superiore. Justine Cassell, direttrice dello Human-Compuer Interaction Institute alla Carnegie Mellon University, afferma:

> From the inside [...] Wikipedia may feel like a fight to get one's voice heard. One gets a sense of this insider view from looking at the "talk page" of many articles, which rather than seeming like collaborations around the construction of knowledge, are full of descriptions of "edit-warring" — where successive editors try to cancel each others' contributions out — and bitter, contentious arguments about the accuracy of conflicting points of view. [...] A woman who wishes to share what she knows with others may not want bitter altercation and successive edit wars. (Cassell, 2011)

La causa principale dunque non starebbe nel fatto che la comunità di Wikipedia ha di per sé delle politiche che sfavoriscono il contributo da parte delle donne, ma piuttosto che alcuni tratti del maschilismo strisciante nei paesi occidentali possono rispecchiarsi nella modalità di collaborazione online e determinare un *gap* per gli utenti donne:

> It is still the case in American society that debate, contention, and vigorous defense of one's position is often still seen as a male stance, and women's use of these speech styles can call forth negative evaluations. Women may be negatively judged for speaking their mind in clear ways and defending their position. A woman who wishes to collaboratively construct knowledge and share it with others might not choose to do

[39] URL: http://www.nytimes.com/roomfordebate/2011/02/02/where-are-the-women-in-wikipedia

so as part of a forum where engaging in debate and deleting others' words is key. (Cassell, 2011)

1.3.4 Volatilità

Quando si parla della "volatilità" di Wikipedia si intende che il contenuto dell'enciclopedia non avrà mai una versione definitiva, ma sarà sempre un *work in progress* fluido e magmatico. C'è il pericolo di non poter di fatto utilizzare Wikipedia come fonte di riferimento, in quanto una qualsiasi citazione fatta in una certa data potrebbe non essere più valida successivamente. D'altro canto citare una versione "storica" di Wikipedia per eliminare il pericolo di inconsistenza dovuto alle modifiche successive significherebbe perdere tutti gli arricchimenti e le correzioni, ovvero, per così dire, "gettare via il bambino insieme all'acqua sporca". Come afferma Peter Denning:

> Contributions and corrections may be negated by future contributors. One of the co-authors of this column found it disconcerting that he had the power to independently alter the Wikipedia article about himself and negate the others' opinions. Volatility creates a conundrum for citations: Should you cite the version of the article that you read (meaning that those who follow your link may miss corrections and other improvements), or the latest version (which may differ significantly from the article you saw)? (Denning et al., 2005)

1.4 Conclusioni

Wikipedia viene giudicata dai suoi critici come una risorsa inaffidabile, lacunosa, soggetta a *bias* di vario genere. Non tutte le critiche mosse al progetto sono tuttavia rilevanti ai fini di questa tesi, che non intende dimostrare la qualità di Wikipedia come enciclopedia, ma la sua adeguatezza come strumento a supporto della classificazione dei documenti sul Web.

In particolare, le accuse che riguardano la sua mancanza di profondità, il suo anti-elitarismo (cioè la tendenza a rifiutare il contributo degli esperti), l'ineleganza dello stile, l'inesattezza delle informazioni non sono qui prese in considerazione, perché sono caratteristiche che non pregiudicano il ruolo di Wikipedia quale specchio del livello della conoscenza media degli utenti del Web. I con-

tenuti dell'enciclopedia sono il frutto di una selezione artificiale di ciò che gli utenti di Internet ritengono possa rappresentarli a livello di consapevolezza e di interpretazione del mondo. Da questo punto di vista Wikipedia è davvero la cosmopedia di Lévy, il corpo documentale pulsante e vivo della Rete, non scevro dalle incongruenze e imperfezioni della Rete stessa. Wikipedia non vuole esser qualcosa di più elevato dei suoi lettori, ma il prodotto del loro stesso livello di conoscenza. Perciò è perfettamente adatta a classificare i documenti in Rete, un'attività che non serve ad altri se non agli internauti stessi. La classificazione sul Web è un processo autoreferenziale, che ha regole solo interne: se si usasse uno strumento con una semantica più profonda o con una forma più raffinata, il compito (sia per classificatori automatici che umani) sarebbe più difficile in quanto il classificatore dovrebbe costantemente tener conto di un *gap* strutturale tra i profili di classificazione (le voci di Wikipedia) e i documenti target (le pagine Web da classificare)[40].

Classificare un documento in base al suo argomento significa assegnargli una o più categorie concettuali (come "Romanticismo", "Friedrich Schiller", "Germania") tra un certo numero di categorie possibili. Poiché questa assegnaziona viene eseguita in virtù della somiglianza tra la categoria concettuale e il documento, il processo di classificazione funziona tanto più correttamente quanto più, per descrivere un determinato argomento, l'autore del documento usa strutture semantiche e linguistiche simili a quelle presenti nella categoria concettuale. L'omogeneità tra le categorie (le voci) di Wikipedia e i documenti (le pagine) presenti sul Web è perciò desiderabile, anche quando questo significhi duplicare in luoghi diversi lo stesso errore o la stessa forma sgraziata.

È certamente vero, come rileva Cohen (2008), che Wikipedia può perpetuare luoghi comuni e pregiudizi, perché è il deposito di una conoscenza popolare consolidata, non lo strumento per scoprire nuova conoscenza. Un esempio molto recente lo si è avuto durante la campagna elettorale delle elezioni politiche italiane del febbraio 2013. Si tratta del "caso Giannino", che Wikipedia italiana descrive in questo modo:

[40] Per il funzionamento del processo di classificazione semantica dei documenti, si veda il paragrafo 2.3.

> Il 18 febbraio 2013, a pochi giorni dalle elezioni, l'economista Luigi Zingales, fondatore con Giannino di Fare, lascia il partito affermando che il giornalista avrebbe mentito sulle proprie credenziali accademiche. In particolare su un master che, secondo alcuni curricula e secondo sue dirette affermazioni il giornalista avrebbe ottenuto alla Booth School of Business di Chicago, ma che non risulta effettivamente mai conseguito. Giannino ha confermato di non detenere né il master né le due lauree attribuitegli, spiegando la vicenda come un equivoco. Il 20 febbraio 2013 annuncia le sue dimissioni da presidente di Fare a favore di Silvia Enrico, ma rimane il "candidato premier" (formalmente Capo della forza politica) a causa dell'impossibilità, data la legge elettorale, di ritirare la propria candidatura.[41]

È esattamente così che tutti noi ricordiamo la vicenda. Tuttavia, al contrario dell'opinione pubblica e dei mass media italiani, Wikipedia aveva fatto "scoppiare" il caso Giannino molto prima del febbraio del 2013, addirittura nell'ottobre del 2011 all'interno della pagina di discussione della voce italiana dedicata a Oscar Giannino[42]. Nel 2011 i wikipediani avevano cercato di verificare le informazioni relative al percorso di studi di Giannino e, trovando alcune incongruenze relative al conseguimento del master a Chicago, avevano contattato direttamente la Booth School of Business per chiedere chiarimenti. La mail di risposta della segretaria della Booth School, Amy Wright, era stata chiara e puntuale:

> Dear O.,
>
> I checked our database and we have no record of attendance or degree conferral by a person named Oscar Giannino. I checked with our Executive Education (non-degree) Office to see if he took a course through them, but again, they have no record of a person named Oscar Giannino.
>
> Kind regards,
>
> Amy Wright[43]

La comunità di Wikipedia si trovava così di fronte a un "conflitto di policy". Da una parte lo scopo di un'enciclopedia è quello di riportare affermazioni veritiere e di non dichiarare il falso, dall'altra Wikipedia prescrive di citare solo fonti "ufficiali", accertate e accertabili su Internet (No Original Research). Sebbene si

[41] URL: http://it.wikipedia.org/wiki/Oscar_Giannino
[42] URL: http://it.wikipedia.org/wiki/Discussione:Oscar_Giannino
[43] URL: http://it.wikipedia.org/wiki/Discussione:Oscar_Giannino#laurea

trattasse di un'informazione probabilmente vera, quella del mancato conseguimento del master da parte di Giannino era il risultato di una ricerca "originale" condotta da un wikipediano e non pubblicata in altri luoghi se non su Wikipedia. Era una conoscenza che non faceva ancora parte del "bagaglio conoscitivo" di Internet, perciò di fatto inutilizzabile. Sono esemplari le parole utilizzate nella pagina di discussione dai wikipediani per descrivere l'*impasse*:

> -Scusate, ma come si concilia il divieto di ricerche originali con l'aver cassato due fonti come l'Istituto Bruno Leoni e il Rotary sulla base di una presunta mail non verificata? Mi sembra una scelta quantomeno eterodossa.
>
> -A me sembra il classico caso di terra piatta vs terra rotonda: magari ha pure ragione Oriettaxx [l'utente che ha pubblicato la mail della Wright], ma se gli addetti ai lavori sono convinti che la terra è piatta, Wikipedia dovrebbe dire che la terra è piatta. O almeno così mi pare si sia sempre fatto.
>
> -Sentite, io rimetto il diploma alla Booth fontato [collegato a una fonte] con l'Istituto Bruno Leoni e le parole dello stesso Giannino. Quando uscirà una fonte secondaria enciclopedica che smentisce ufficialmente Giannino ne riparleremo. Nel frattempo siete pregati di smetterla di fare i giornalisti d'inchiesta e tornare a compilare un'enciclopedia.
>
> -Una fonte deve essere esterna, se se lo dice lui da solo, non vale come fonte su wiki. (Indipendentemente, non è che un semplice diploma sia un'informazione così enciclopedica, quindi io toglierei indipendentemente dal fontato o meno. Wikipedia deve essere una lista di diplomi ottenuti?)[44]

Questo genere di scelte sono dunque estremamente consapevoli in Wikipedia. La comunità sa bene qual è l'attività che sta svolgendo e il modo in cui la sta svolgendo. Anche se i critici possono non ritenerlo un principio valido, per Wikipedia tramandare l'eredità culturale consolidata sul Web è prioritario rispetto allo scoprire la verità o al correggere false credenze.

Una critica invece da considerare rilevante per la classificazione è quella relativa alla copertura diseguale degli ambiti del sapere umano. La diversa copertura in Wikipedia si manifesta in DBpedia con una maggiore o minore presenza di entità e di relazioni tra le entità. Questo di ripercuote sui software di classificazione basati su DBpedia, perché i testi su alcuni argomenti saranno classificati con

[44] Ibidem.

maggiore precisione rispetto ad altri. Nel caso specifico di TellMeFirst, infatti, i documenti di argomento scientifico e tecnologico danno risultati migliori rispetto a quelli di ambito umanistico. La maggiore cura e ricchezza di dettagli presente nelle voci di carattere tecnico-scientifico in Wikipedia da origine a profili delle classi più complessi e accurati in questi settori. Volendo descrivere il processo di classificazione di TellMeFirst nella maniera più semplice e sintetica possibile, si può dire che esso consiste nel confronto tra un documento target (il testo da classificare) e i profili delle classi del sistema, ovvero l'insieme di tutte le voci di Wikipedia. I profili valutati come i più simili al documento target vanno a costituire le categorie sotto cui il documento deve essere classificato, nella forma di URI di DBpedia. Per costruire il profilo della classe "Alessandro Manzoni", per esempio, TellMeFirst raccoglie tutte le informazioni presenti in Wikipedia relative al concetto di Alessandro Manzoni, sia dalla voce http://en.wikipedia.org/wiki/Alessandro_Manzoni sia dalle altre voci dell'enciclopedia in cui compare il link a http://en.wikipedia.org/wiki/Alessandro_Manzoni, sia infine dagli *infobox* che compaiono sulla destra di ogni voce (collezionati poi da DBpedia). Quanto più corposa e ricca di informazioni è la voce su Alessandro Manzoni in Wikipedia, quanto più numerosi sono i link ad Alessandro Manzoni in altre pagine, e quanto più consistente è l'*infobox* che ne struttura i dati puntuali (luogo e anno di nascita, occupazione, genere letterario, ecc), tanto più complesso sarà il profilo della categoria "Alessandro Manzoni" nel sistema di classificazione di TellMeFirst. Gli articoli di Wikipedia che hanno un profilo di categoria molto complesso hanno in TellMeFirst anche una maggiore possibilità di essere considerati argomento di un testo. Per questo, la disomogeneità delle voci dell'enciclopedia può riflettersi nella qualità della classificazione in sistemi automatizzati come TellMeFirst, favorendo gli argomenti di carattere tecnologico, informatico, scientifico (biologia, medicina, chimica) e riguardanti lo spettacolo o gli hobby (cinema, televisione, giochi, sport), che nel corpus documentale abbiamo visto avere maggiore approfondimento, rispetto a quelli degli altri settori del sapere. Anche la maggiore consistenza della versione inglese di Wikipedia rispetto a quella italiana comporta una disomogeneità nella classificazione: i testi in lingua inglese, che utilizzano Wiki-

pedia e DBpedia inglesi come *training set*[45], danno risultati di classificazione migliori in termini di precisione rispetto ai testi in italiano.

Un altro problema importante è quello che prende le mosse dal giudizio di McHenry sulla produzione collaborativa della conoscenza e che si collega col concetto di "volatilità". Wikipedia è un costante *work in progress* prodotto e gestito dalla comunità degli utenti di Internet: ma in ogni singolo momento essa può rappresentare adeguatamente la conoscenza di questa comunità? Se blocchiamo il processo di generazione del sapere in un istante di tempo x (come accade quando utilizziamo Wikipedia come *training set* per un classificatore automatico), chi ci garantisce che il corpus delle voci nell'istante x sia adeguato come lo era nell'istante x-n dove n è una quantità arbitraria di tempo? Può accadere che nell'intervallo di tempo che va da x-n a x alcune voci di Wikipedia siano rimosse a seguito di un atto di vandalismo, per poi essere riportate alla versione precedente (con la funzione di *rollback*[46]) nell'istante x+n. Per quanto piccolo possa essere il valore di n, esisterà sempre un intervallo di tempo non nullo da x a x+n dove la conoscenza contenuta in Wikipedia non rappresenta realmente la conoscenza media della comunità. Tenendo in conto il principio di "eventualismo" che è alla base di Wikipedia come di tutti i progetti *open source*, si può dire che Wikipedia "tende a un tempo infinito" ad essere rappresentativa della conoscenza della comunità, ma che in ogni singolo istante di tempo ci sono possibilità che non lo sia. Poiché i casi di vandalismo si presentano con una frequenza non troppo alta e la correzione degli errori avviene molto rapidamente (anche grazie alla "legge di Linus"[47]), è comunque altamente probabile che in media la conoscenza veicolata da Wikipedia rispecchi ciò che la comunità ha stabilito essere il proprio bagaglio conoscitivo.

---

[45] Per il significato di *training set* si veda il paragrafo 2.3. La dimensione di Wikipedia inglese è circa 6 volte superiore a quella di Wikipedia italiana. Il *dump* XML compresso di Wikipedia inglese in data 02/01/2013 è di 9,7 GB. Quello di Wikipedia italiana in data 09/01/2013 è di 1,7 GB. URL: http://dumps.wikimedia.your.org/backup-index.html

[46] URL: http://en.wikipedia.org/wiki/Help:Reverting

[47] Si veda il paragrafo 1.1.

Capitolo 2

DBpedia, Linked Open Data e classificazione semantica dei contenuti

2.1 Web Semantico e Linked Open Data

Il Word Wide Web è basato, oggi come alla sua nascita nel 1991, su HTML, un linguaggio che descrive come le informazioni devono essere disposte e visualizzate su una pagina online per essere fruite dagli esseri umani. Infatti il Web si è sviluppato come un medium per l'esposizione dell'informazione direttamente agli utenti finali, senza che vi fosse una reale enfasi sulla possibilità per le macchine di comprendere ed elaborare in background i documenti pubblicati (Alesso et al., 2009; Ryan, 2010). Dato che il WWW ha origine dall'idea di ipertesto, Internet è oggi caratterizzato da contenuti prevalentemente testuali accompagnati da aggiunte grafiche o audiovisive. I browser sono gli strumenti applicativi che permettono di trasformare queste informazioni in qualcosa di comprensibile agli umani, nonché di spostarsi da una risorsa Web all'altra per mezzo di collegamenti ipertestuali (link).

Il limite strutturale del Web sta nella ricerca dell'informazione. Non essendo contenuta nelle pagine HTML alcuna indicazione esplicita riguardante il loro significato, lo strato semantico del Web deve essere costruito sopra quello esistente, come un "soprastrato" delle risorse informative. I motori di ricerca hanno immediatamente avuto un grande successo sul Web proprio perché sopperivano alla carenza semantica del WWW fornendolo di un *layer* semantico "implicito", costruito in maniera algoritmica a partire dal contenuto dei testi e dalle ricerche degli utenti. Nonostante il significato delle pagine Internet non sia specificato,

Google è in grado di ricondurre un certo numero di documenti (mediamente nell'ordine delle centinaia di migliaia) alla nostra esigenza di ricerca. Digitiamo "arte barocca" e Google ci fornisce istantaneamente risultati che sembrano avere a che fare con l'arte barocca, secondo un criterio che è oscuro all'utente comune, benché il funzionamento dell'algoritmo, chiamato PageRank (Page et al., 1998) sia noto nelle sue linee più generali agli esperti del settore. Spetta all'utente scegliere la parola chiave che gli sembra più efficace per una particolare ricerca, adeguandosi ai meccanismi interni del motore di ricerca: per esempio ognuno sa che la punteggiatura non conta in una query su Google, o che, se si usa un termine ambiguo, conviene aggiungere qualche altra parola per non ottenere risultati inattesi. Quando il bisogno informativo è più complesso, l'utente sa di dovere digitare un numero maggiore di query e di cercare poi all'interno dei documenti trovati ciò che gli occorre.

Questo meccanismo, rivolto all'utente umano, ha già di per sé qualcosa di problematico. Non pochi studiosi hanno ravvisato il pericolo di affidare completamente a un motore di ricerca proprietario l'appagamento di un nostro bisogno informativo (Maurer et al., 2007; Carr, 2010; Morozov, 2011). In primo luogo il criterio di ranking dei risultati nella maggior parte dei motori di ricerca ha come componente fondamentale la popolarità della risorsa Web piuttosto che la pertinenza alla query o l'attendibilità della fonte. Fa notare per esempio Sara Monaci:

> La logica della "popolarità" è sovrana per il motore di ricerca: gli indici elaborati dal PageRank svolgono un ruolo fondamentale nel rendere visibile e significativo ciò che si annida tra le trame della Rete. […] Tale valutazione rimane tuttavia estranea al contenuto specifico del documento: si fonda infatti non tanto su una valutazione di tipo semantico quanto di una valutazione per così dire "sintattica" che elabora non tanto il significato o i significati di una risorsa quanto il suo essere collegata ad altre risorse simili o affini per argomento. (Monaci, 2008, p. 74)

Un'altra criticità importante è quella relativa alla "personalizzazione" dei motori di ricerca come Google, che rischia di dare origine al fenomeno noto come *filter bubbles* (Pariser, 2011). È una situazione in cui l'algoritmo di un sito Web filtra in maniera selettiva quali sono le informazioni utili per l'utente in base alle informazioni relative all'utente stesso (come la sua posizione, i click precedenti,

gli ultimi acquisti, ecc). A partire dal 2009 Google utilizza di default la ricerca personalizzata, separando gli utenti a loro insaputa da una grossa parte dell'informazione presente sul Web, in particolare dalle fonti più distanti a livello di interessi e gusti personali. In questo modo essi hanno una minore esposizione a punti di vista differenti e rischiano di rimanere isolati intellettualmente nella propria "bolla informativa".

> With Google personalized for everyone, the query "stem cells" might produce diametrically opposed results for scientists who support stem cell research and activists who oppose it. "Proof of climate change" might turn up different results for an environmental activist and an oil company executive. In polls, a huge majority of us assume search engines are unbiased. But that may be just because they're increasingly biased to share our own views. More and more, your computer monitor is a kind of one-way mirror, reflecting your own interests while algorithmic observers watch what you click. (Pariser, 2011, p. 5)

Ancora maggiori sono i problemi relativi all'attività in Rete degli agenti software. L'infrastruttura del Web rende difficoltosi i processi di acquisizione ed elaborazione automatica della conoscenza, in quanto non esiste una maniera unica di rappresentare l'informazione (il significato dei documenti e dei servizi, la loro fonte, il loro grado di attendibilità, la loro funzione). Ogni servizio sul Web ha la propria interfaccia (Application Programming Interface, API) e risponde con le proprie strutture di dati. Se è vero che gli sviluppatori possono maneggiare separatamente tali servizi e comporli per ottenere risposte a domande complesse, lo stesso non si può dire per le macchine, che non trovano l'interoperabilità necessaria per poter agire correttamente.

Obiettivo del Semantic Web (o Web dei Dati) è descrivere il significato dell'informazione pubblicata sul Web per consentirne il reperimento sulla base di una comprensione precisa della sua semantica. Il Semantic Web aggiunge struttura alle risorse accessibili online in maniera che non siano soltanto fruibili dagli esseri umani, ma anche processate rapidamente dagli agenti software.

> A new Web architecture called the Semantic Web offers users the ability to work on shared knowledge by constructiong new meaningful representation on the Web. Semantic Web research has developed from the tradition of Artificial Intelligence (AI)

> and ontology languages and offers automated –processing through machine-understandable metadata. (Alesso et al., 2009, p. 15)

Invece che come testo (che non può essere elaborato da un computer senza l'ausilio di algoritmi di Natural Language Pocessing), l'informazione sul Web Semantico è pubblicata come (meta)dato strutturato per mezzo di un linguaggio molto semplice chiamato RDF (Resource Description Framework)[48]. Ciò implica che l'unità fondamentale del Web non è più il documento, ma il singolo dato, nella forma di una proposizione logica che unisce un soggetto ad un oggetto attraverso un predicato. L'insieme di questi fatti puntuali può descrivere oggetti reali ("Napoleone" "è morto a" "Sant'Elena") oppure risorse informative sul Web, come documenti o contenuti multimediali ("La pagina Web X" "ha come argomento" "Napoleone"). Ogni cosa o risorsa informativa online ("entità" nel seguito) è identificata nel Web dei Dati per mezzo di un riferimento univoco (URI[49]). Dataset diversi possono avere riferimenti diversi alla stessa entità, ma è buona pratica che indichino esplicitamente la reciproca identità degli URI per mezzo di predicati speciali come per esempio *owl:sameAs*.[50] È quindi possibile per un agente software navigare tra le entità presenti sul Web alla ricerca di fatti, anche tra dataset esposti da *publisher* diversi.

> Linked Data is simply about using the Web to create typed links between data from different sources. These may be as diverse as databases maintained by two organisations in different geographical locations, or simply heterogeneous systems within one organisation that, historically, have not easily interoperated at the data level. Technically, Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets. (Berners-Lee et al., 2009, p. 2)

Si può considerare ogni unità informativa del Web semantico (detta anche "tripla") come corrispondente al record di una database, dove il soggetto è l'identificativo della riga, il predicato l'identificativo della colonna e l'oggetto il valore del campo. Considerando i dataset esposti sul Web Semantico come basi

[48] URL: http://www.w3.org/RDF/
[49] URL: http://www.w3.org/Addressing/#background
[50] URL: http://www.w3.org/TR/owl-ref/#sameAs-def

di dati liberamente accessibili sul Web, l'intero spazio informativo può essere definito Web dei Dati. Per reperire le informazioni puntuali da un dataset sul Web dei Dati è necessario utilizzare un linguaggio di interrogazione (*query language*) simile a SQL: è stato dunque introdotto il linguaggio SPARQL (Simple Protocol and RDF Query Language[51]), per mezzo del quale si possono eseguire query direttamente sull'*endpoint* di un dataset che espone triple RDF.

> An RDF store is similar to a relational database or an XML store. Not surprisingly, in the early days of RDF, a number of different query languages were available, each supported by some RDF-based product or open-source project. From the common features of these query languages, the W3C has undertaken the process of standardizing an RDF query language called SPARQL. (Allemang et al., 2008, p. 67)

Una caratteristica fondamentale del Web Semantico è la classificazione delle entità. Le classi sono espresse anch'esse da URI e legate alle entità per mezzo dello speciale predicato *rdf:type*[52], che indica una relazione di appartenenza ("Giovanni" "appartiene alla classe" "Persona"). Le classi e i predicati utilizzati all'interno di un dataset possono essere formalmente descritti da un'ontologia. Il termine ontologia viene introdotto in campo informatico da Tom Gruber nel paper *A Translation Approach to Portable Ontology Specifications* (1993). Gruber pensa all'applicazione delle ontologie nel campo dell'intelligenza artificiale (nello specifico, nella rappresentazione della conoscenza) e le definisce come «la specificazione esplicita di una concettualizzazione»:

> An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what "exists" is that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, in the context of AI, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g.,

[51] URL: http://www.w3.org/TR/rdf-sparql-query/
[52] URL: http://www.w3.org/TR/rdf-schema/#ch_type

> classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. (Gruber, 1993, p. 2)

Come in filosofia l'ontologia è la materia che studia l'essere e le sue forme, in IA un'ontologia descrive formalmente, con il linguaggio della logica, "tutto ciò che esiste" in un dominio di conoscenza più o meno esteso. Le ontologie servono ai sistemi intelligenti per conoscere il mondo esterno e per poter agire in esso. La rappresentazione della conoscenza in forma *machine-readable* serve in primo luogo alle macchine per prendere la decisione opportuna sulla base degli elementi conosciuti (che possono essere dati ambientali, fatti storici, variabili economiche, ecc). Nel Web Semantico un'ontologia contiene tutte le classi presenti all'interno di un dataset e può organizzarle secondo una tassonomia per mezzo del predicato *rdfs:subClassOf*[53]. Le classi sono considerate come insiemi di entità e possono essere descritte per mezzo dei predicati che queste entità possiedono. Per esempio posso descrivere la classe "Scrittore" come il sottoinsieme delle entità della classe "Persona" che sono legate all'entità "opera letteraria" dal predicato "essere autore di". Per scrivere ontologie sul Web si utilizza lo specifico linguaggio OWL (Web Ontology Language[54]), basato sulla logica descrittiva (Description Logic) e serializzato in RDF/XML.

A seconda dei loro confini tematici (il loro *scope*) le ontologie possono essere di tre tipi:

1) Ontologie di dominio. Si tratta di schemi concettuali che rappresentano la conoscenza di un dominio applicativo più o meno specifico, come la biologia, la geografia di una nazione o una raccolta di monete.
2) Ontologie linguistiche. Modellano e mettono in relazioni i termini di una lingua attraverso le definizioni, le radici, le figure retoriche, i sinonimi e i contrari, ecc, Se multilingua, associano parole che hanno il medesimo significato in due o più idiomi diversi. La più nota ontologia linguistica di-

[53] URL: http://www.w3.org/TR/rdf-schema/#ch_subclassof
[54] URL: http://www.w3.org/TR/owl-ref/

sponibile in Rete è WordNet[55], con oltre 90.000 corrispondenze lessicali in lingua inglese.

3) Upper-level ontologies. Un'ontologia di alto livello rappresenta il tentativo di modellare l'intera realtà in un unica tassonomia, da cui le specifiche ontologie ereditano al fine di classificare i diversi ambiti del sapere. Hanno un livello di astrazione molto elevato e sono spesso frutto di competenze multidisciplinari che coniugano le tecnologie dell'informazione con la filosofia analitica e l'epistemologia. Una delle più note è DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering[56]) , risultato dell'attività dell'Istituto per le Scienze e le Tecnologie Cognitive del CNR di Trento.

"Linked Data" si riferisce ad un modo di pubblicare ed interconnettere dati strutturati sul Web usando il linguaggio RDF. La nozione di "dati collegati" era presente già nella prima teorizzazione del Web Semantico (Berners-Lee at al., 2001), ma solo in seguito è entrata *in auge* in ambito informatico fino a sovrapporsi e a sostituire quella di Semantic Web. In realtà, più precisamente, i Linked Data sono una parte del disegno del Web Semantico, quella più basilare che riguarda i "mattoncini" senza i quali il Web Semantico non può essere costruito. Il disegno del Web Semantico è molto più vasto, interessa per esempio anche gli agenti intelligenti che utilizzano questi dati, le ontologie che devono essere progettate per integrare la semantica dei diversi dataset e il *reasoning* che può essere effettuato sui linguaggi di rappresentazione della conoscenza. Infatti il paper di esordio sul Web Semantico, scritto nel 2001 da Tim Berners-Lee, James Hendler e Ora Lassila, dal titolo *The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*, è un testo rivoluzionario, profetico, che propone scenari tecnologici ancora oggi considerati futuribili.

Gli autori delineano un progetto di Internet e della società basato sulle tecnologie del Semantic Web. I compiti più complessi di reperimento dell'informazione sono delegati alle macchine, agenti intelligenti che operano sul Web interrogando servizi ed interagendo tra loro per acquisire ed elaborare dati. Nell'esempio con-

[55] URL: http://wordnet.princeton.edu/

[56] URL: http://www.loa.istc.cnr.it/DOLCE.html

tenuto nel paper, un uomo chiamato Pete affida al suo *smart agent* il compito di trovare uno specialista in fisioterapia per la madre e di fissare un ciclo di trattamenti. I criteri di selezione di Pete sono allo stesso tempo vaghi e complessi, proprio come accade in una situazione reale: il fisioterapista dovrebbe essere affidabile, qualificato, non lontano da casa della mamma e compatibile con la sua assicurazione; inoltre gli appuntamenti dovrebbero tener conto della disponibilità di Pete stesso, perché sarà lui ad accompagnare la madre alle visite.

> At the doctor's office, Pete instructed his Semantic Web agent through his handheld Web browser. The agent promptly retrieved information about Mom's prescribed treatment from the doctor's agent, looked up several lists of providers, and checked for the ones in-plan for Mom's insurance within a 20- mile radius of her home and with a ratingof excellentor very goodon trusted rating services. It then began trying to and a match between available appointment times (supplied by the agents of individual providers through their Web sites) and Pete's and Lucy's busy schedules. (Berners-Lee et al., 2001, p. 2)

Il compito degli agenti è agevolato dal fatto che l'informazione sul Web Semantico è espressa attraverso linguaggi comuni e decidibili, che condividono la stessa struttura di significati e permettono ai software di eseguire operazioni di *reasoning* per risolvere i problemi più complessi. Il terreno linguistico per l'interoperabilità delle macchine sul Web è dato dalle ontologie, rappresentazioni formali dei vari domini di conoscenza, basate sulla logica descrittiva.

I concetti su cui insistono particolarmente gli autori del paper sono dunque legati al Knowledge Representation and Reasoning (KRR) e all'intelligenza artificiale:

> The real power of the Semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information and exchange the results with other programs. The effectiveness of such software agents will increase exponentially as more machine-readable Web content and automated services (including other agents) become available. (Berners-Lee et al., 2001, p. 5)

Circa otto anni dopo, nella conferenza TED del 2009 intitolata *Tim Berners-Lee on the next Web*[57], il focus dell'autore non è più sugli stessi concetti. È invece incentrato in maniera evidente sulla tematica degli Open Data e sulla necessità per le istituzioni, gli enti di ricerca, ma anche per i cittadini stessi di rilasciare i propri dati sul Web in modo da renderli liberamente accessibili. Provando a sottoporre a TagCroud[58] il paper del 2001 e la trascrizione della conferenza TED del 2009, si ottengono i seguenti risultati (vedi Figura 1 e 2).

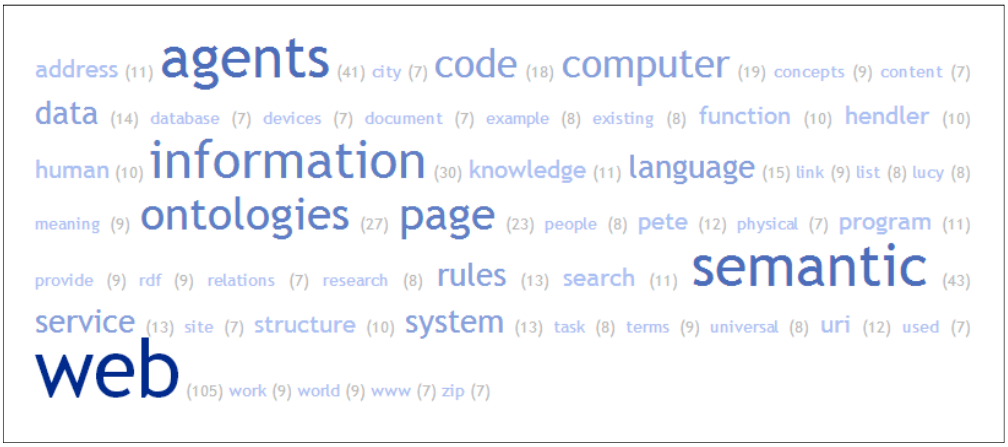


Figura 1 - Tag cloud del paper di Tim Berners-Lee *The Semantic Web*, 2001

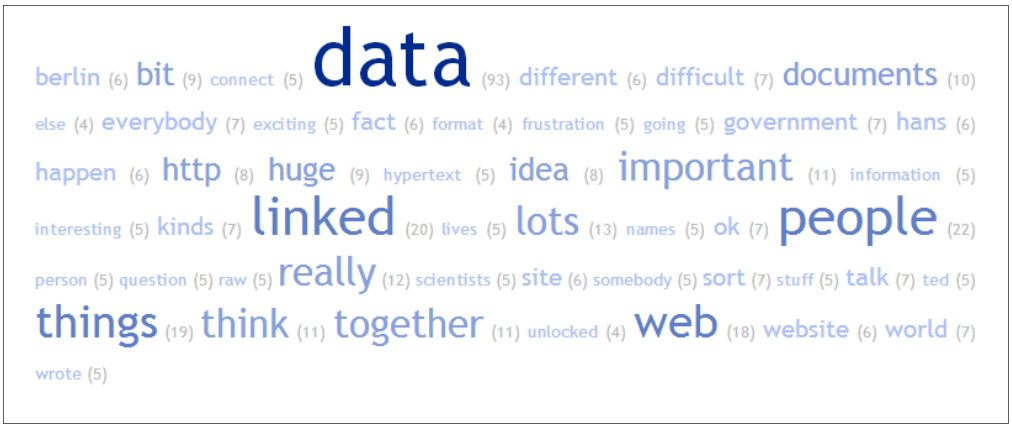


Figura 2 - Tag cloud della conferenza TED di Tim Berners-Lee nel 2009

[57] URL: http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html
[58] URL: http://tagcrowd.com/

Alla prevalenza delle parole “agente”, “semantica”, “ontologia”, “linguaggio”, “codice”, ”computer”, “regola” (nel senso di *inference rule*) nel primo testo, corrisponde una netta predominanza delle parole “dati” e “dati collegati” nel secondo, seguite da concetti non tecnici come “persona”, “documento” e “Web”.

Il Semantic Web, secondo la visione di Tim Berners-Lee et al. (2001), avrebbe impiegato un linguaggio standard per la rappresentazione della conoscenza contenuta nelle pagine Web e per la dichiarazione delle regole in base alle quali gli agenti software dovevano utilizzare questa conoscenza. Questa informazione strutturata poteva essere utilizzata in primo luogo per migliorare l'accuratezza delle ricerche sul Web, in quanto un motore di ricerca avrebbe cercato solo all'interno di pagine contenenti il markup di un concetto preciso piuttosto che affidarsi a keyword vaghe ed ambigue. Inoltre applicazioni più complesse avrebbero potuto collegare questi dati strutturati ad ontologie che contenevano regole d'inferenza per dedurre nuova informazione a partire da quella già nota. Questo meccanismo prende il nome di *reasoning* ed è consentito dalla decidibilità di OWL: il Web Ontology Language deriva la sua semantica dalla famiglia delle logiche descrittive, quindi ha a disposizione tutti gli assiomi e le regole di questi modelli logici. L'articolo originario di Tim Berners-Lee si concentra dunque principalmente su come aggiungere un significato *machine-readable* all'informazione già pubblicata nelle pagine Web. Di conseguenza, nei primi anni Duemila la ricerca ha avuto tra i principali filoni lo sviluppo di nuove ontologie per il Web, in diversi domini come la biologia (Gene Ontology[59]), il diritto (LKIF Core Ontology of Basic Legal Concepts[60]), la geografia (W3C Geospatial Ontologies[61]), i beni culturali (CIDOC Conceptual Reference Model[62]). Oltre alle ontologie di dominio, grande importanza è stata data anche alle ontologie fondative (o *upper ontologies*), come CommonKADS (Schreiber et al., 2000), SUMO (Pease et al., 2002), DOLCE (Gangemi et al., 2002), ecc. Questi modelli descrivono concetti di livello molto generale per gettare le basi, le fondamenta appunto, della conoscenza umana a cui dovrebbero fare riferimento tutte le diverse ontologie di dominio.

[59] URL: http://www.geneontology.org/

[60] URL: http://www.estrellaproject.org/lkif-core/

[61] URL: http://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/

[62] URL: http://www.cidoc-crm.org/

Le ontologie di dominio possono ereditare da una o più classi di un'ontologia fondativa, utilizzandola come uno strumento per l'interoperabilità semantica. DOLCE per esempio, specifica i concetti basilari di "endurante" e "perdurante": l'endurante è una entità che mantiene nel tempo le caratteristiche fondamentali del suo essere (una persona, per esempio, è persona dalla sua nascita fin dopo la sua morte); il "perdurante" invece è diverso in diversi istanti di tempo (come per esempio un evento o un arco temporale). A partire da due concetti così astratti, vi è l'intera struttura delle classi dell'ontologia, a cui si possono agganciare alberature specifiche per qualsiasi dominio.

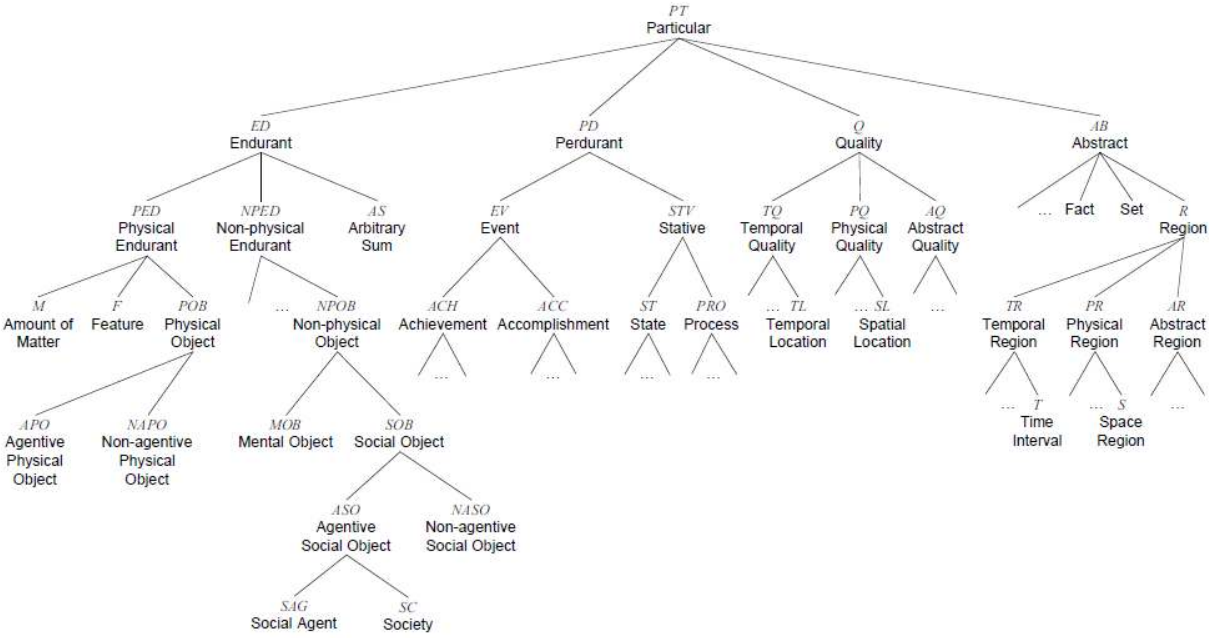


Figura 3 - Categorie di base in DOLCE (Masolo et al., 2003, p. 14)

La ricerca sulle ontologie fondative si è accompagnata allo studio di metodologie per progettare e implementare ontologie coerenti e funzionali. Questo campo prende il nome di *ontology engineering* e concilia competenze specifiche nel dominio descritto dall'ontologia (medicina, diritto, ecc) a competenze informatiche e logico-insiemistiche. Uno dei risultati più notevoli in questo settore è stato il metodo OntoClean (Guarino et al., 2002), un approccio per validare l'adeguatezza ontologica delle relazioni tassonomiche. Essa si basa su nozioni astratte che traggono origine dall'ontologia analitica (Runggaldier et al., 1998), come essenza

e identità, e che vengono utilizzate per caratterizzare gli aspetti rilevanti del significato delle classi e delle proprietà che compongono un'ontologia. Questi aspetti sono rappresentati attraverso metaproprietà formali (*rigidity*, *identity* e *unity*) che impongono alcuni vincoli sulla struttura tassonomica: l'analisi di tali vincoli aiuta a valutare e convalidare le scelte ontologiche effettuate. I quattro vincoli fondamentali sono sintetizzati da Guarino in questo elenco:

> 1) If q is anti-rigid, then p must be anti-rigid;
> 2) If q carries an identity criterion, then p must carry the same criterion;
> 3) If q carries a unity criterion, then p must carry the same criterion;
> 4) If q has anti-unity, then p must also have anti-unity.
>
> (Guarino et al., 2004, p. 6)

È sulla base dei principi di OntoClean che è stata progettata a partire dal 2001 l'ontologia fondazionale DOLCE: oltre a rispettare le restrizioni metodologiche sopra descritte, DOLCE compie alcune scelte concettuali precise, che la rendono un'ontologia "del senso comune", "moltiplicativa" e "rigida" (Cairo et al., 2012). Contemporaneamente all'avanzamento dell'ontology engineering, nei primi anni del Duemila si è assistito alla nascita di numerosi software per l'*editing* grafico di ontologie, come per esempio Protégé[63] e HOZO[64] e per il *reasoning* automatico su RDFS e OWL, come Racer (Haarslev et al., 2001) e Pellet (Sirin et al., 2007).

Verso la fine del decennio, tuttavia, ci si è resi conto che la ricerca sull'*ontology engineering* e sulle tematiche annesse stava di fatto rallentando il processo di diffusione del Semantic Web nel mondo reale. Costruire ontologie, per di più secondo precise regole di *good design*, era complesso e dispendioso: sembrava quindi un investimento abbastanza azzardato, visto che i dati a disposizione in RDF sul Web erano ancora assai pochi. Questa difficoltà è stata definita da Hart et al. (2013) «the cold-start problem»: come convincere i data publisher ad aggiungere markup in RDF o in OWL alle proprie pagine Web con la sola prospettiva a lungo termine di poter ottenere benefici quando "tutti gli altri

[63] URL: http://protege.stanford.edu/
[64] URL: http://www.hozo.jp/

publisher" avessero adottato lo stesso markup? Il problema più immediato da risolvere non era stabilire quale pattern di *ontology design* avrebbe garantito un *reasoning* più efficace, ma liberare i dati dai *silos* in cui erano conservati e cominciare a pubblicarli sul Web in forma aperta, indipendentemente dal legame che essi potessero avere con risorse informative preesistenti sul Web.

> It was perhaps the concentration of the Semantic Web research community on developing large ontologies and optimizing ever-more-complex reasoners that led some to rethink the direction in which the Semantic Web was heading. The Linked Data movement started with a return to the drawing board to concentrate on exposing data to the Web that was hidden in proprietary databases, structured in myriad ways. To do this, they recommended structuring data in a standard format: RDF, which could also be used to specify links in to and out of each dataset. (Hart et al., 2013, p. 14)

Il cambio di contesto concettuale ravvisato nelle *tag cloud* del paper di Tim Berners-Lee del 2001 e del suo discorso al TED nel 2009 rispecchia dunque l'evoluzione del dibattito intorno al Web Semantico avvenuta nel primo decennio del Duemila. L'attenzione rivolta ai linguaggi di rappresentazione e *all'ontology engineering*, che ha dato il via a progetti di straordinaria portata ma difficilmente realizzabili come DOLCE, ha caratterizzato la prima parte del decennio. In seguito tuttavia il focus si è gradualmente spostato verso la questione dei dati aperti, con le problematiche politiche e giuridiche derivate, collegandosi con gli analoghi movimenti Open Content e Open Access.

Senza i dati aperti, infatti, non è pensabile cominciare a realizzare il disegno del Web Semantico. Gli Open Data sono la materia prima del Web Semantico, la benzina senza cui il motore non può funzionare. Il passo indietro di Tim Berners-Lee e della comunità era necessario per non far rimanere i Linked Data semplicemente un sogno fantascientifico. Nel 2007 è nata ufficialmente la comunità dei Linked Open Data[65] (nome che rispecchia la compresenza nel movimento delle due istanze "linked" e "open"), che ha trovato linfa vitale nelle iniziative di Open Government del 2009 Data.gov[66] negli USA e Data.gov.uk[67] in Inghilterra, a cui

[65] URL: http://linkeddata.org/
[66] URL: http://www.data.gov/
[67] URL: http://data.gov.uk/

sono seguiti i portali Open Data di molte altri stati. Per promuovere l'adozione dei Linked Data da parte delle pubbliche amministrazioni, il W3C ha messo on-line specifiche linee guida finalizzate a incoraggiare la pubblicazione di dati pubblici e a diffondere le *best practices* che le autorità pubbliche possono seguire. In aggiunta a queste linee guida, Tim Berners-Lee ha pubblicato un documento di natura più divulgativa in cui spiega perché le tecnologie Linked Data sono il modo migliore per rispondere alle tre esigenze per le quali i dati delle amministrazioni pubbliche sono messi on-line: «to increase accountability, contribute valuable information about the world, and to enable government, the country, and the world to function more efficiently» (Berners-Lee, 2009). Gli Open Linked Data hanno trovato particolare diffusone in ambito governativo, dove le necessità di trasparenza, di interoperabilità e di partecipazione dal basso erano più sentite:

> Negli ultimi due anni le istanze dei movimenti per l'accesso aperto alla conoscenza si sono rivolte anche all'"informazione del settore pubblico" (PSI). Incoraggiato dai cambiamenti in atto e dai risultati ottenuti, un nuovo movimento dal basso, conosciuto con il nome di Open Government Data, si sta diffondendo nei paesi industrializzati con l'obiettivo di ottenere l'accesso libero e proattivo ai dati di un ambito specifico: quello delle istituzioni politiche e della pubblica amministrazione. I dati, affermano gli aderenti al movimento, devono essere liberi da limitazioni tecnologiche e legali che ne impediscano il riuso, la modifica e la combinazione con altri dati, così da far accedere alle informazioni in maniera molto diretta e trasparente, per renderci cittadini più consapevoli e dunque più liberi. (Di Donato, 2010)

Il passo avanti compiuto da Tim Berners-Lee nel 2001 era troppo grande, in quanto la sua visione del Web e della società richiedeva sforzi eccessivi da parte degli *information publisher* per arrivare ad una semantica condivisa per l'integrazione di piattaforme e servizi. In aggiunta, il disegno generale poteva apparire rischioso, perché presupponeva una totale fiducia nell'operato delle macchine, i cui meccanismi di *reasoning* sui dati e di verifica delle fonti non potevano essere sotto il pieno controllo dell'utente. Al contrario, identificando negli Open Data il presupposto fondativo del Web Semantico, il movimento nel suo complesso ha trovato basi teoriche più semplici ed è riuscito a connettersi ad obiettivi

fondamentali della società dell'informazione come la trasparenza governativa e la condivisione del sapere.

2.2 DBpedia: il punto di riferimento per il Web dei Dati

DBpedia è un progetto nato nel 2007 dalla collaborazione di Università Libera di Berlino, Università di Lipsia e OpenLink Software[68]. È finalizzato all'estrazione di dati strutturati dal corpus di Wikipedia e alla pubblicazione dei dati stessi sul Web in formato Linked Data. La knowledge base (KB) estratta dalla versione inglese di Wikipedia descrive più di 3,77 milioni di entità, tra cui 764.000 persone, 573.000 luoghi, 333.000 opere creative, 72.000 film e 45.000 aziende. Ci sono versioni di DBpedia in 111 lingue: in totale esse hanno 8 milioni di collegamenti a immagini, 24.4 milioni di link a pagine Web esterne e 27.2 milioni di collegamenti a dataset in RDF esterni (Bizer et al., 2009).

Le voci di Wikipedia sono costituite per lo più da testo libero, ma possono anche contenere informazioni semi-strutturate secondo il modello editoriale del Wiki. Tali informazioni comprendono i cosiddetti *infoboxes*, ovvero gli schemi che si trovano sulla destra di molte pagine e che riassumono alcuni dati fondamentali per l'argomento trattato (Figura 8) , le categorie di appartenenza (per esempio "Astronomia" o "Nati nel 1928"), le immagini, le coordinate geografiche, i link a pagine Web esterne, le pagine di disambiguazione (Figura 6), i reindirizzamenti tra le pagine (per esempio da "USA" a "United States of America") e i collegamenti a edizioni di Wikipedia in lingua diversa. Il progetto DBpedia estrae queste informazioni strutturate o semi-strutturate da Wikipedia e le trasforma in una vasta KB. Per compiere questa operazione il gruppo di sviluppo ha implementato un Extraction Framework[69] in grado di prelevare le diverse categorie di informazioni, parsificarle, serializzarle e ripubblicarle sul Web. L'output è costituito da dataset statici in sintassi N-Ttriples[70] (.nt) e Turtle[71] (.ttl) scaricabili dal

[68] URL: http://www.openlinksw.com/
[69] URL: https://github.com/dbpedia/extraction-framework
[70] URL: http://www.w3.org/2001/sw/RDFCore/ntriples/

portale del progetto[72], da un endpoint SPARQL esposto attraverso OpenLink Virtuoso[73] e da Linked Data liberamente accessibili per mezzo di un browser HTTP (vedi architettura in Figura 4).

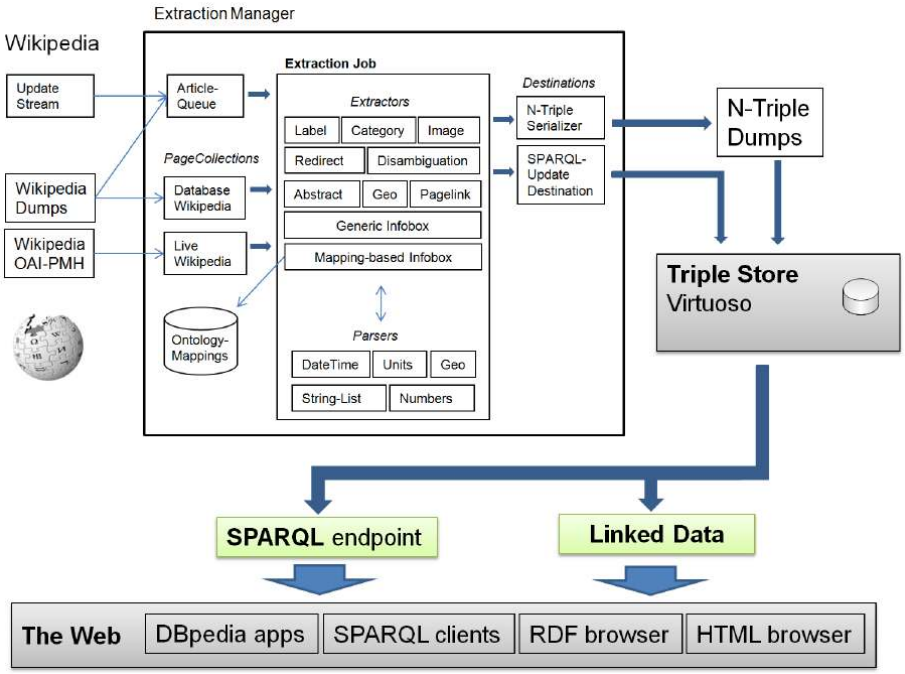


Figura 4 - Architettura del DBpedia Extraction Framework (Bizer et al., 2009)

I dataset estratti dal Framework sono elencati di seguito.

1. *Labels*. I titoli delle voci di Wikipedia, che sono utilizzati come stringhe nella proprietà *rdfs:label* della corrispettiva risorsa di DBpedia.
2. *Abstracts*. Gli abstract sono di due tipi: il primo, più breve, contiene la frase d'apertura dell'articolo di Wikipedia ed è rappresentato con la proprietà *rdfs:comment*; il secondo, più esteso, contiene tutto il testo che precede l'indice dei contenuti dell'articolo ed è rappresentato attraverso un'apposita proprietà *dbpedia:abstract*.
3. *Interlanguage links*. Sono i link che connettono gli articoli sullo stesso argomento nelle molteplici versioni linguistiche di Wikipedia: in questo

[71] URL: http://www.w3.org/TeamSubmission/turtle/
[72] URL: http://wiki.dbpedia.org/Downloads38
[73] URL: http://virtuoso.openlinksw.com/

modo ogni risorsa di DBpedia è collegata alle risorse equivalenti nelle altre lingue.

4. *Images*. Gli URL di una o più immagini di Wikimedia Commons[74] utilizzate all'interno della pagina di Wikipedia. Sono rappresentate attraverso le proprietà *foaf:depiction* e *dbpedia:thumbnail*.
5. *Redirects*. I reindirizzamenti tra i titoli di Wikipedia collezionati da DBpedia per consentire ad altre applicazioni di costruire catene di transitività che conducono alla risorsa giusta.
6. *Disambiguation*. Le pagine di disambiguazione di Wikipedia sono raccolte da DBpedia per consentire di scoprire i diversi significati di termini omonimi (come "Donatello" nell'esempio di Figura 6). La proprietà usata per la loro rappresentazione è *dbpedia:disambiguates*.
7. *External links*. Sono i link a pagine Web esterne a Wikipedia; vengono rappresentati per mezzo della proprietà *dbpedia:reference*.
8. *Pagelinks*. I link che collegano internamente le pagine di Wikipedia sono rappresentati con la proprietà *dbpedia:wikilink*.
9. *Homepages*. Si tratta degli URL dei siti Web di entità come aziende, organizzazioni, ecc. Sono rappresentati attraverso la proprietà *foaf:homepage*.
10. *Categories*. Le categorie sono il mezzo che Wikipedia ha a disposizione per classificare le proprie voci in macroargomenti (come "Italian philosophers", o "Romantic poets", ecc.). Ogni categoria diventa in DBpedia uno *skos:concept* e le relazioni tra le categorie vengono rappresentate attraverso la proprietà *skos:broader*[75].
11. *Geo-coordinates*. DBpedia estrae le coordinate di tutti i luoghi presenti in Wikipedia e le rappresenta in formato WGS84[76].

Il passaggio più complesso nell'intero processo di produzione della KB di DBpedia consiste nell'estrazione delle informazioni dagli *infobox*. Wikipedia ha specifiche norme per la creazione degli *infobox*, formalizzate in linee guida relative ai *template* che ogni diverso tipo di *infobox* deve seguire (c'è un *template* per i luoghi, uno per gli animali, per le persone, per gli artisti ecc.). Il problema è che non soltanto le diverse versioni linguistiche di DBpedia hanno regolamentato gli

[74] URL: http://commons.wikimedia.org/wiki/Main_Page
[75] Su SKOS, si veda: http://www.w3.org/2004/02/skos/
[76] URL: http://it.wikipedia.org/wiki/WGS84

infobox in maniera diversa, ma all'interno di una stessa versione gli utenti non sempre hanno rispettato le linee guida. Ne risulta che gli *infobox* hanno spesso tra loro una diversa struttura, utilizzano stringhe diverse per gli attributi che si riferiscono allo stesso concetto (per esempio "birth place" e "place of birth") e hanno come valori talvolta semplici stringhe, talvolta link ad altre risorse di Wikipedia (*wikilinks*). Per affrontare questo disordine, il progetto DBepdia ha usato due diversi tipi di strategia, senza far prevalere l'una sull'altra:

1) Generic Infobox Extraction. Le informazioni contenute in un *infobox* vengono trasformate in una serie di triple in cui il soggetto è costituito dall'URI della risorsa, il predicato dal *namespace* http://dbpedia.org/property/ concatenato al nome dell'attributo e l'oggetto dal valore dell'attributo. Sebbene vengano utilizzati alcuni algoritmi per cercare di ricondurre l'oggetto ad un URI di DBpedia o a un valore *literal*[77] (e di inferire il *datatype*[78] nel caso di un *literal*), il procedimento presenta un margine di errore piuttosto alto. Inoltre non vengono ricondotti ad una stessa proprietà gli attributi sinominici, perché non omografi. Questa strategia consente perciò di ottenere la massima copertura delle proprietà degli *infobox* a scapito delle precisione.
2) Mapping-based Infobox Extraction. Si mappano gli attributi degli *infobox* su un'ontologia che raccoglie sotto un'unica proprietà gli attributi sinonimici. Questa ontologia è stata creata manualmente e consiste di 720 proprietà che mappano 2350 attributi. Il mapping definisce regole puntuali su come parsificare i valori degli attribuiti, specificando se si tratta di un URI o di un *literal* e indicandone il *datatype*. Lo svantaggio di questo approccio è che copre solo 350 *template* degli *infobox*, fornendo dunque dati su circa 843.000 entità, un numero molto esiguo se comparato alle 1.462.000 entità coperte dalla Generic Infobox Extraction. Questa strategia consente di ottenere la massima accuratezza nella rappresentazione delle proprietà degli *infobox* a scapito delle copertura.

A partire dalla release 3.8, DBpedia pubblica alcuni dataset specificamente pensati per il supporto al Natural Language Processing, chiamati appunto DBpe-

[77] URL: http://www.w3.org/TR/rdf-concepts/#section-Literals
[78] URL: http://www.w3.org/TR/REC-rdf-syntax/#section-Syntax-datatyped-literals

dia NLP Datasets[79]. Questi sono stati prodotti all'interno del progetto DBpedia Spotlight[80] e messi a disposizione per il download nel portale di DBpedia. Si tratta di DBpedia Lexicalizations Dataset, DBpedia Topic Signatures, DBpedia Thematic Concepts e DBpedia People's Grammatical Genders.

DBpedia Lexicalizations Dataset è un insieme di triple in formato N-Triples che mappano gli URI di DBpedia alle diverse *surface form* che essi hanno in Wikipedia. Le *surface form* vengono raccolte sia dai titoli delle voci di Wikipedia (dopo un semplice *parsing* che elimina per esempio le parentesi di disambiguazione) sia dagli *anchor text* dei *wikilink*. In tal modo, si ottengono tutte le possibili forme alternative in cui compare una risorsa DBpedia all'interno di Wikipedia: ad esempio la risorsa http://dbpedia.org/resource/United_States è associata alle *surface form* "Stati Uniti", "Stati Uniti d'America", "USA", "America", "US", ecc.

DBpedia Topic Signatures è un TSV che collega ogni risorsa di DBpedia a un certo numero di *token* che rappresentano la sua *bag of words*, ovvero le parole che meglio descrivono il suo contesto semantico. Vengono raccolti i paragrafi in cui occorrono i *wikilink*, poi sono tokenizzati e trasformati in un vettore pesato in cui il peso di ogni *token* è stabilito con la formula TF-IDF. Il vettore viene ordinato in base al peso dei *token*, quindi vengono estratti i primi n elementi e connessi alla risorsa DBpedia. Un esempio di Topic Signature per la risorsa http://dbpedia.org/resource/Apple_Records (i primi tre *token*) è "beatles album released".

DBpedia Thematic Concepts è un insieme di triple in formato N-Triples che raccolgono gli URI di DBpedia per argomento. Le categorie di DBpedia sono collegate alle varie risorse attraverso la proprietà *dcterms:subject*, dunque un modo per raccogliere tutte le entità che appartengono a una stessa categoria è via query SPARQL. Una maniera alternativa è quella di ottenere tutti i *wikilink* contenuti nelle pagine di Wikipedia dedicate a una specifica categoria, come per esempio Astronomia (http://en.wikipedia.org/wiki/Category:Astronomy). Queste liste vengono inserite nel DBpedia Thematic Concepts Dataset.

[79] URL: http://wiki.dbpedia.org/Datasets/NLP?v=10lj
[80] URL: https://github.com/dbpedia-spotlight/dbpedia-spotlight

DBpedia People's Grammatical Genders è un N-Triples che associa a ogni risorsa di DBpedia appartenente alla classe dbpedia:Person il suo genere, nella forma ":Male" o ":Female". Per ottenere l'informazione sul genere, viene parsificata la corrispettiva pagina di Wikipedia alla ricerca di pronomi personali, aggettivi possessivi, ecc. che hanno forma diversa per il maschile e per il femminile (come "he", "she", "her", "him", "himself", "herself"). Dalla frequenza di queste forme viene inferito il genere della risorsa con un certo grado di probabilità.

Per consentire agli utenti di scoprire nuova informazione a partire da DBpedia, la KB ha al suo interno molti collegamenti ad altri dataset della Linked Data Cloud. Da DBpedia partono circa 4.9 milioni di link verso l'esterno, che esprimono l'identità semantica delle risorse di DBpedia con le risorse di Amsterdam Museum, BBC Wildlife Finder, RDF Bookmashup, Bricklink, CORDIS, DailyMed, DBLP, DBTune, Diseasome, DrugBank, EUNIS, Eurostat, CIA World Factbook, Flickr Wrappr, Freebase, GADM, GeoNames, GeoSpecies, Project Gutenberg, Italian Public Schools, LinkedGeoData, LinkedMDB, MusicBrainz, New York Times, OpenCyc, OpenEI, Revyu, SIDER, RDF-TCM, UMBEL, US Census, WikiCompany, WordNet Classes, YAGO2 e GHO (vedi la distribuzione dei link in Figura 5). Questa identità è espressa attraverso la relazione *owl:sameAs*.

Oltre ad avere molti *outgoing links* (collegamenti che puntano verso l'esterno), Dbpedia possiede anche il più alto numero di *ingoing links* dai dataset della Linked Data Cloud (collegamenti che partono da un dataset esterno e arrivano a DBpedia). Essendo una pratica comune quella di collegare le entità del proprio dataset alle entità di DBpedia, per via della sua importanza e completezza, DBpedia è diventata lo snodo principale del Web dei Dati, l'*hub* da cui passano gran parte delle relazioni tra i diversi dataset. Basta uno sguardo alla nota immagine della Cloud nel settembre del 2011 (Figura 7), per rendersi conto della centralità di DBpedia nel contesto del Semantic Web. Lo spazio informativo di DBpedia è in questo modo enormemente ampliato: gli sviluppatori possono costruire interessanti e utili *mashup* che combinano dati provenienti da diverse fonti, e gli agenti software (come i browser e i *crawler* semantici) sono in grado di

navigare il Web dei Dati nella modalità *follow your nose*[81] senza soluzione di continuità.

Data Source	No. of Links	Data Source	No. of Links
Freebase	2,400,000	WikiCompany	25,000
flickr wrappr	1,950,000	MusicBrainz	23,000
WordNet	330,000	Book Mashup	7,000
GeoNames	85,000	Project Gutenberg	2,500
OpenCyc	60,000	DBLP Bibliography	200
UMBEL	20,000	CIA World Factbook	200
Bio2RDF	25,000	EuroStat	200

Figura 5 – Distribuzione degli *outgoing link* da DBpedia verso gli altri dataset

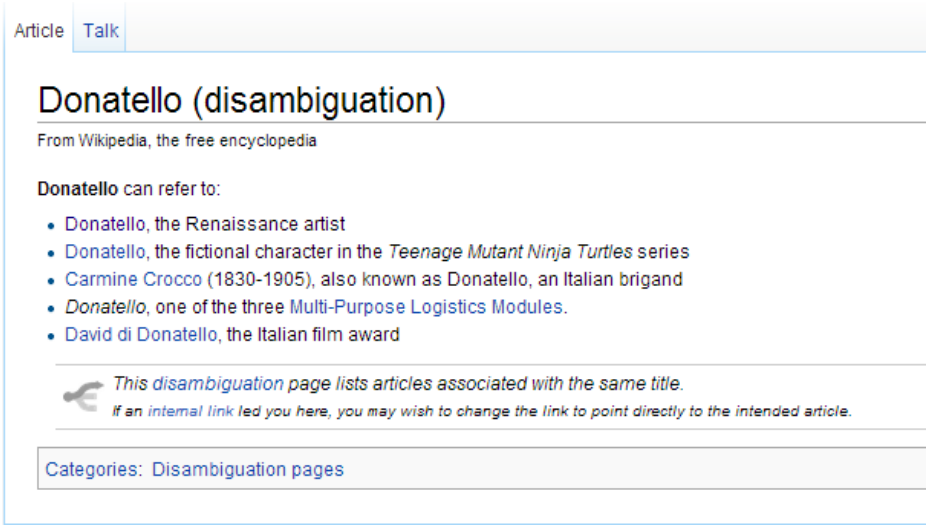


Figura 6 - Pagina di disambiguazione per il titolo “Donatello” in Wikipedia inglese

[81] URL: http://www.w3.org/2001/sw/wiki/Linking_patterns#.E2.80.9CFollow_your_nose.E2.80.9D

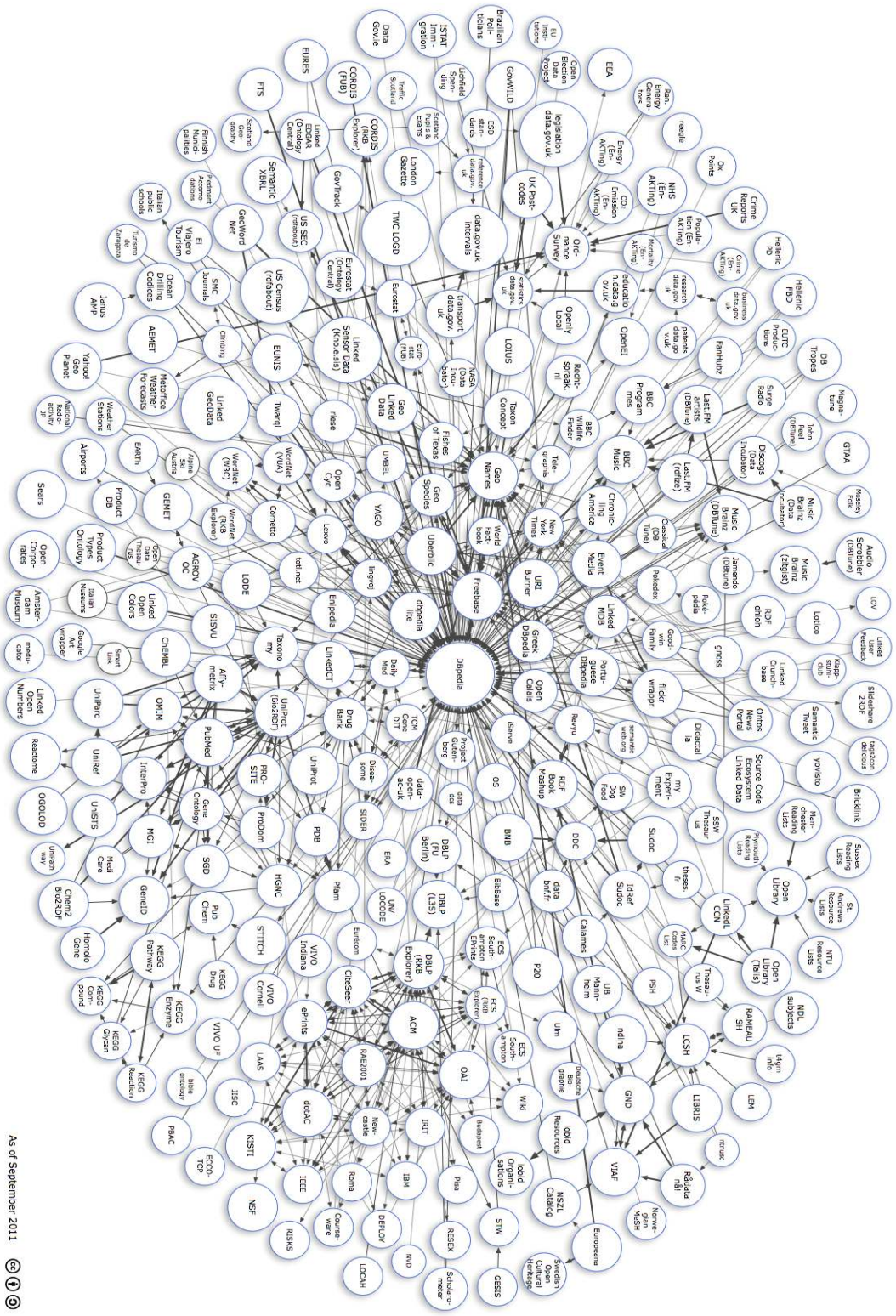
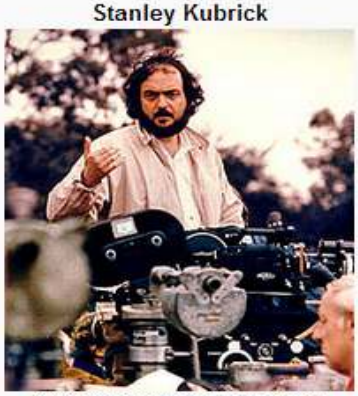


Figura 7 - Linked Data cloud al settembre del 2011[82]

[82] URL: http://lod-cloud.net/

Stanley Kubrick



Kubrick filming *Barry Lyndon* (1975)

Born	July 26, 1928 Manhattan, New York, United States
Died	March 7, 1999 (aged 70) Harpenden, Hertfordshire, England, United Kingdom
Cause of death	Heart attack
Nationality	American
Ethnicity	Jewish
Occupation	Film director, producer, screenwriter, cinematographer, editor
Years active	1951–1999
Notable work(s)	*2001: A Space Odyssey*, *Dr. Strangelove*, *The Shining*, *A Clockwork Orange*, *Barry Lyndon*, *Full Metal Jacket*
Influenced by	Max Ophüls, Sergei Eisenstein, Elia Kazan, G. W. Pabst, Vsevolod Pudovkin, Orson Welles
Influenced	Woody Allen, Paul Thomas Anderson, Tim Burton, James Cameron, Coen brothers, Frank Darabont, Guillermo del Toro, Todd Field, David Fincher, Terry Gilliam, David Lynch, Michael Mann, Gaspar Noé, Christopher Nolan, Nicolas Winding Refn, George A. Romero, Martin Scorsese, Ridley Scott, Steven Spielberg, Quentin Tarantino, Lars von Trier
Spouse(s)	Toba Etta Metz (1948–51; divorced) Ruth Sobotka (1954–57; divorced) Christiane Harlan (1958–99; his death)

Figura 8 - Infobox della pagina dedicata a Stanley Kubick di Wikipedia inglese[83]

[83] URL: http://en.wikipedia.org/wiki/Stanley_Kubrick

Wikidata[84] potrebbe essere realmente *the big next thing* nel giovane campo del Web dei Dati. Si tratta di un progetto ancora in fase di sviluppo, nato in seno alla fondazione Wikimedia nell'aprile del 2012 e finanziato da tre potenti *investor* come Allen Institute for Artificial Intelligence[85], Gordon and Betty Moore Foundation[86] e Google. Wikidata mira alla creazione e alla pubblicazione sul Web di un gigantesco database comprensivo di tutte le conoscenze strutturate che l'umanità possa raccogliere. Il meccanismo è simile a quello di Wikipedia: saranno gli stessi utenti della piattaforma a inserire, correggere e validare i dati della KB grazie a una interfaccia di tipo modulo (*form*). Si prevede anche la possibilità di specificare la fonte del dato, in maniera da poterne verificare l'affidabilità. I contenuti inseriti in Wikidata, come quelli presenti in Wikipedia, saranno rilasciati sotto licenza libera e perciò riutilizzabili tanto dagli esser umani quanto dai software che raccolgono e processano le informazioni sul Web.

È inevitabile il confronto con DBpedia, anche se la differenza tra i due progetti risulta abbastanza chiara. DBpedia pubblica i propri dati in RDF sul Web, ma, una volta pubblicati, essi sono "statici" ed occorre una modifica a Wikipedia e un'ulteriore estrazione per modificarli. Wikidata invece consente agli utenti di modificare direttamente il database senza dover passare da Wikipedia; anzi, è Wikipedia a poter sfruttare l'informazione presente in Wikidata per creare/modificare gli *infobox* e verificare i contenuti dell'enciclopedia.

Questa differenza tuttavia non può essere considerata sufficiente per una separazione "funzionale" dei due progetti. Nella documentazione di Wikidata (ovviamente in formato Wiki), è presente una pagina[87] dedicata specificamente al confronto tra i due progetti, dove si legge:

> Overlap between the two projects
>
> - Both projects publish URIs for entities based on Wikipedia.
>
> - Both projects publish RDF data about entities. The source of the data is very different: whereas DBpedia extracts the data from the infoboxes, Wikidata will collect

[84] URL: http://meta.wikimedia.org/wiki/Wikidata/en
[85] URL: http://www.ai-squared.org/
[86] URL: http://www.moore.org/
[87] URL: http://meta.wikimedia.org/wiki/Wikidata/Notes/DBpedia_and_Wikidata

> data entered through its interfaces. Data in Wikidata will also be annotated with its provenance: it does not simply state the population of Germany, but it also requires a source to be given for the data. The two data repositories will co-exist. If Wikidata gets established and collects an interesting amount of data, the relationship between the two datasets should be further explored.

Di fatto, se Wikidata realizzerà i suoi obiettivi, si troverà ad avere un vantaggio nei confronti di DBpedia, perché l'approccio collaborativo alla produzione e alla modifica del dato è più efficiente e democratico rispetto a quello di estrazione automatica su base tecnologica. Ci si aspetterebbe che i due progetti siano portati avanti congiuntamente, o dallo stesso gruppo di lavoro, perché il loro fine è il medesimo, sebbene cambi la tecnica per ottenerlo. Ma questo connubio sembra essere più conveniente per DBpedia, che potrebbe fare a meno dell'Extraction Framework e attingere i propri dati da Wikidata in maniera dinamica, mantenendo il dominio "dbpedia" come il punto di riferimento per il Web dei Dati. Wikidata invece non sembra avere particolare bisogno di DBpedia, perché è in grado di implementare tutti gli *step* della filiera dei Linked Data, dalla creazione alla pubblicazione, all'aggiornamento, alla rivalidazione.

Il nuovo progetto è nato in maniera indipendente da DBpedia, come si evince dall'intervento di Chris Bizer, il coordinatore di DBpedia, nella sezione "Potential Contributions from DBpedia" della stessa pagina di documentazione[88]. Il team di Wikidata ha risposto *inline*, perciò si riporta qui la sezione attribuendo le singole frasi a Bizer o a Wikidata:

> Bizer: The DBpedia team is excited that the Wikipedia community is moving towards reorganizing the importance of handling structured data within Wikipedia and we wish the Wikidata project to be a great success. If the Wikidata project sees fit, the project is of course highly invited to reuse whatever part of DBpedia that you think is useful. Potential contributions from DBpedia could for instance be:
>
> the reuse DBpedia source code where ever the Wikidata project sees fit (things don't need to be implemented twice).
>
> Wikidata: Wikidata will survey the DBpedia source code and keep an eye on reusable components.

[88] Ibidem.

Bizer: DBpedia data could be used to bootstrap the Wikidata repository with initial content (the DBpedia team would be happy to help, if this is desired by Wikidata).

Wikidata: The Wikidata team does not decide on the content of the Wikidata site. I am sure the community, once it is established, will have a discussion about this topic.

Bizer: Any other help the Wikidata project is interested in, just send a mail to dbpedia-discussion or dbpedia-developers.

Wikidata: Thank you! We will.

La scelta degli URI da parte di Wikidata sarà probabilmente fondamentale per chiarire come i due progetti potranno coesistere. Da quanto emerge dall'apposita pagina di documentazione[89], la scelta potrebbe ricadere su di uno schema del tipo "http://wikidata.org/id/Q{id}", dove {id} è un codice numerico che contraddistingue una pagina di Wikidata. Il modello degli URI di Wikipedia/DBpedia viene invece considerato pericoloso, perché inserisce una parte di semantica (mutevole) in un identificativo che dovrebbe rimanere stabile nel tempo. In effetti, esempi di cambiamento degli URI nel passaggio da una versione all'altra di DBpedia sono molto comuni.

Che si tratti o meno di una opportunità per DBpedia, indubbiamente Wikidata è una grande opportunità per il Web Semantico. Wikidata può diventare il nucleo attorno a cui si sviluppa la rete dei dati sul Web e quindi la scienza collaborativa, coadiuvando DBpedia nel suo ruolo fondamentale di *hub* tra diversi dataset della Cloud. La cooperazione tra Wikidata e DBpedia è di certo auspicabile, per essere certi che la nuova iniziativa si integri senza intoppi nel più vasto disegno dei Linked Open Data.

[89] URL: http://meta.wikimedia.org/wiki/Wikidata/Essays/URI_scheme

2.3 Annotazione e classificazione semantica del testo

2.3.1 Che cosa si intende per classificazione ed annotazione semantica

La classificazione e l'annotazione semantica sono operazioni che è possibile svolgere sui testi, sia in maniera automatizzata che manuale, utilizzando una KB di riferimento. Prima di approfondire questo argomento, è utile introdurre in sintesi i concetti di classificazione e di annotazione dei testi, che, per quanto simili, presentano caratteristiche e problematicità specifiche.

Per classificazione di un testo si intende la sua assegnazione a una o più classi preesistenti. Dato un insieme di classi distinte, ognuna con un suo profilo e descritta da una serie di caratteristiche, il processo di classificazione determina a quale classe appartiene un documento testuale, ovvero quale classe descrive meglio le caratteristiche del documento. Il criterio di scelta delle caratteristiche rilevanti per la classificazione è fondamentale ed è stabilito *a priori* dal classificatore (umano o software). Per esempio possiamo essere interessati alla lunghezza dei documenti e perciò classificarli a seconda del numero di parole, oppure alla data in cui sono stati prodotti, o ancora alla lingua. Questo genere di classificazione non è però detta "semantica". La classificazione è semantica quando gli elementi di interesse per la classificazione si riferiscono al "significato" del documento. Categorizzare un documento in base al suo argomento principale ("Cina", "Astronomia", "Alessandro Del Piero", ecc) è un esempio di classificazione semantica, ma lo è parimenti suddividere i documenti in base all'argomento meno trattato o in base all'atteggiamento dell'autore verso la materia (positivo, critico, indulgente, obiettivo, ecc)[90].

[90] Questo è il campo della *sentiment analysis*, che cerca di individuare in maniera automatica qual è il punto di vista di chi scrive un'opinione su un determinato argomento sulla base del contesto linguistico utilizzato. Negli ultimi anni la *sentiment analysis* ha trovato grande diffusione in ambito commerciale per raccogliere le opinioni degli utenti su un prodotto (*opinion mining*) o per capire il grado di soddisfazione della clientela nei confronti di un determinato marchio (*brand reputation*).

Annotazione e classificazione semantiche hanno una difficoltà in più rispetto ai loro corrispettivi non semantici. Mentre per classificare un insieme di documenti in base al numero di parole basta assegnare loro un valore numerico, per classificare gli stessi documenti in base al significato non basta assegnare loro una stringa. Le stringhe infatti, come ad esempio i tag che usiamo sul Web, possono avere un certo grado di ambiguità che inficia in maniera più o meno grave l'efficacia della categorizzazione. La stringa "Einaudi" per esempio può indicare entità diverse come "Giulio Einaudi", fondatore della nota casa editrice, "Luigi Einaudi", secondo presidente della Repubblica Italiana, "Ludovico Einaudi", compositore e pianista, ecc. La stringa "onda" può riferirsi a un movimento della superficie delle acque, alla caratteristica di un segnale fisico, a un tratto di un elettrocardiogramma, ecc. Per questo motivo la classificazione semantica ha bisogno di un "dizionario delle classi" che identifichi e descriva formalmente le diverse categorie, in maniera da rendere impossibile qualsiasi ambiguità. Questa specie di dizionario è chiamato knowledge base del classificatore semantico, o anche ontologia di riferimento. La KB ha al proprio interno un modo per indicare univocamente le classi: utilizza, al posto delle stringhe, codici identificativi, come accade nelle basi di dati, chiamati ID, UID o URI, ecc.

Vi sono fondamentalmente tre macro-metodologie per classificare i documenti (Manning et al., 2008): quella manuale, quella basata su regole (*rule-based*) e quella basata su apprendimento (*machine learning-based*). Il metodo manuale è ovviamente il più tradizionale ed è nato molto prima della nascita del computer. Per secoli i libri nelle biblioteche sono stati accuratamente classificati da esseri umani, esperti in tale dominio, per essere organizzati e cercati più facilmente. Questa pratica è ancora oggi la più diffusa sia in ambito cartaceo che digitale. La maggior parte dei CMS per la pubblicazione di giornali online o di blog consentono ai propri autori di classificare o taggare i propri articoli e post. Anche i social network più conosciuti come Twitter e Facebook permettono di etichettare i contenuti con parole chiave. La strategia manuale può riguardare anche la classificazione semantica. Vi sono prodotti commerciali che consentono una classifica-

zione semantica via Wikipedia/DBpedia in modo manuale: tra i più noti c'è Faviki[91], nato nel 2010. Si legge nel loro sito:

> Free-word tags do not have defined meanings, so it isn't always clear what a particular tag represents. Does the tag "jaguar" represent the animal, the car company, or the operating system? Faviki uses Common tags - unique, well-defined concepts from Wikipedia that allow you to state what a web page is exactly about, letting your computer understand you better.[92]

La seconda macro-strategia, quella basata su regole, prevede l'uso di specifiche query per raccogliere i documenti di interesse. Queste query definiscono una "regola di classificazione" per la quale o un documento è interessante secondo il criterio scelto o non lo è. Un esempio può essere "tutti i documenti il cui autore è X e che contengono la parola Y più di 10 volte, sono classificati nella categoria C". Una persona tecnicamente qualificata (ad esempio un esperto di dominio bravo a scrivere le espressioni regolari) può creare set di regole che possono competere o superare l'accuratezza dei classificatori automatici basati su apprendimento, tuttavia può essere difficile trovare qualcuno con questa competenza specialistica e dunque il procedimento è considerato dispendioso.

La metodologia basata su apprendimento è invece quella che sta ottenendo maggiori consensi all'interno della comunità scientifica per la sua accuratezza e scalabilità. Si tratta di un insieme di tecniche che utilizzano il Machine Learning per consentire a un software di classificare automaticamente un nuovo documento sulla base della classificazione che documenti simili ad esso hanno ricevuto in precedenza. Il sistema "impara" a classificare i documenti da esempi di classificazione che gli sono stati sottoposti in passato (il cosiddetto *training set*). L'intervento umano nella classificazione non è del tutto eliminato, in quanto i documenti del *training set* sono etichettati manualmente (per questo tale procedimento si dice "supervisionato"[93]), tuttavia l'attività di *labeling* è molto meno

[91] URL: http://www.faviki.com

[92] URL: http://www.faviki.com/pages/welcome/

[93] Esistono anche approcci non supervisionati al machine learning, in cui le classi non sono stabilite a priori e non vengono forniti al sistema esempi pre-annotati, tuttavia danno mediamente risultati meno precisi (Manning et al., 2008).

dispendiosa di quella di scrittura delle regole. Spesso infatti si hanno già esempi di testi etichettati, per esempio paper scientifici, articoli di giornale pre-taggati e corpora costruiti ad hoc per il Machine Learning.

La classificazione *machine learning-based* può utilizzare diversi tipi di algoritmo: tra i più noti ricordiamo K-Nearest Neighbors, Naive Bayes, Support Vector Machines, Hidden Markov Model, Decision Tree. Tuttavia gli step per eseguire la classificazione sono i medesimi per tutti i sistemi basati su Machine Learning (Basili et al., 2005), ovvero:

1) *Corpus processing*. Comprende tutte le operazioni di filtraggio e formattazione dei testi utilizzati come *training set*, in modo da renderli *machine readable* e omogenei a livello di encoding, struttura interna, ecc.
2) *Extraction of relevant information*. Il training corpus è ripulito dalle componenti inutili per la classificazione, come per esempio le *stop words*, una lista di parole di uso comune nella lingua di riferimento che hanno una distribuzione pressoché simile in ogni classe.
3) *Normalization*. La normalizzazione comprende operazioni di lemmatizzazione e/o di *stemming*. Nella lemmatizzazione le categorie morfologiche complesse vengono ridotte alla loro forma base, nello *stemming* si estrae soltanto la radice dei termini per approssimare il concetto sottostante (per esempio la radice "acquis-" approssima il concetto alla base dei termini "acquisire", "acquisizione", "acquisto", ecc).
4) *Feature Selection*. Le *feature* di un documento sono gli elementi che servono a rappresentarlo all'interno dello spazio informativo. *Feature* dei documenti testuali sono normalmente considerate singole parole o lemmi, mentre per esempio i pixel possono essere le *feature* di un'immagine digitale. La *feature selection* è il processo in base a cui si sceglie un sottoinsieme di termini che occorrono nel *training set* e si utilizza solo quel sottoinsieme per rappresentare il documento stesso. Diminuendo il numero di termini da tenere in considerazione nel processo di classificazione, si rende questo processo più veloce ed efficiente. Inoltre si possono eliminare le *feature* che contribuiscono a creare "rumore" nella classificazione (*noise features*): se per esempio la parola "claustrofobia" accidentalmente ricorresse parecchie volte in alcuni documenti che trattano l'argomento "Italia", il sistema potrebbe essere erroneamente indotto a considerare

"claustrofobia" come un elemento importante per la classe Italia. Con la *feature selection* invece si selezionano solo gli elementi davvero rilevanti, scartando tutto il resto.

5) *Feature Weighting*. Tra gli elementi del documento rilevanti per la classificazione, alcuni di essi hanno un "peso" maggiore di altri. I nomi propri di luogo, per esempio, possono essere particolarmente importanti per una classificazione dei testi su base geografica. Dunque a ogni *feature* di un documento del *training set* viene assegnato un diverso peso che ne indica l'importanza in vista della classificazione. Esistono vari modelli di *feature weighting*: uno dei più noti ed utilizzati nell'ambito dell'Information Retrieval è TF-IDF (Term Frequency-Inverse Document Frequency). In base a questo algoritmo, il peso di un termine all'interno di un documento è direttamente proporzionale alla frequenza del termine nel documento e inversamente proporzionale alla frequenza del termine nell'intero corpus. La logica alla base di questa scelta è che parole molto comuni devono avere un peso minore nella classificazione rispetto parole meno diffuse, perché queste ultime possono meglio rappresentare la specificità linguistica e concettuale del singolo documento. Per esempio la parola "anno" che appare 3 volte in un documento ha meno peso della parola "decadentismo" che compare lo stesso numero di volte, perché "anno" ha una diffusione molto più ampia all'interno dell'intero corpus documentale.
6) *Similarity estimation*. Questa operazione consiste nel trovare un coefficiente di "somiglianza" tra due o più documenti in base alle *feature* pesate che vengono estratte da ogni documento. La *similarity estimation* è la chiave di volta del processo di classificazione, perché un nuovo documento è associato a una o più classi preesistenti proprio in base alla similarità del profilo di questo documento col profilo delle classi.
7) *Application of a Machine Learning model*. Una volta scelto un criterio per determinare la similarità tra due o più documenti, è necessario raccogliere manualmente un insieme di documenti pre-classificati per creare il "profilo" di ciascuna classe. I diversi modelli di classificazione *machine learning-based* differiscono soprattutto nel modo in cui, a partire dai documenti pre-classificati, si giunge al profilo delle classi. Nel caso di *feature* pesate con valori reali (utilizzando per esempio un numero reale compreso tra 0 e 1 per rappresentare il peso di una parola), si può pensare all'insieme dei documenti del *training set* come ad uno spazio continuo

dove le classi sono delle "linee di confine" tra i documenti (Figura 9). Ogni documento è perciò rappresentato come un vettore pesato di *feature* all'interno di uno spazio vettoriale n-dimensionale, dove n è il numero totale di *feature*. Gli algoritmi di Machine Learning per la classificazione che utilizza il Vector Space Model (VSM) possono essere visti come il tentativo di determinare nella maniera più efficace e meno dispendiosa i confini tra le diverse classi.

8) *Inference*. Il passo successivo alla scelta del modello di Machine Learning è l'utilizzo di tale modello per la classificazione di nuovi documenti. La similarità (o la funzione di appartenenza) tra i nuovi documenti e i profili delle classi è utilizzata per inferire una scelta di classificazione.

9) *Testing*. L'accuratezza del classificatore è valutata facendo uso di un insieme di documenti pre-etichettati diverso dal *training set* chiamato *gold standard*. Le etichette prodotte dal classificatore sono quindi confrontate con quelle del *gold standard* per ottenere un valore numerico che indica la distanza (il margine di errore) tra la classificazione umana e quella automatica.

Figura 9 - Rappresentazione bidimensionale di tre classi di documenti all'interno di uno spazio vettoriale. Pallini, crocette e quadratini sono documenti, mentre le linee continue sono i confini tra le classi (Manning et al., 2007, p. 290)

Per annotazione di un testo si intende la pratica comune di aggiungere informazione sul testo stesso attraverso sottolineature, note, commenti, tag o link. Come è chiaro, si tratta di un'attività tradizionalmente manuale che ha acquista-

to forme di automatizzazione con l'emergere delle tecnologie del Natural Language Processing e del Text Mining. Anche l'annotazione dei documenti, come la classificazione, può essere semantica. Si parla di annotazione semantica quando sul testo di un documento vengono aggiunte informazioni riguardanti il suo significato o il significato dei singoli elementi che la compongono. Per fare ciò si utilizzano principalmente link che collegano una parola, un'espressione o un sintagma ad una risorsa informativa sul Web o ad una entità presente in una KB. Esempi manuali di annotazione semantica con risorse provenienti dal Web sono all'ordine del giorno nei siti Internet, nei blog e nei giornali online. Gli autori di contenuti sul Web creano collegamenti ipertestuali non solo per consentire al lettore di approfondire un certo argomento o di attingere l'informazione da un'altra fonte, ma anche per disambiguare un termine o un'espressione, indicando in maniera esplicita il riferimento a un certo concetto anziché a un altro. La frase "Apple è stata fondata nel 1968" può essere per esempio annotata come "Apple è stata fondata nel 1968" per specificare che si sta parlando dell'etichetta discografica creata dai Beatles e non dell'azienda di prodotti elettronici. Questo utilizzo dell'annotazione semantica è quello che maggiormente ci interessa nel contesto tecnologico. I software di annotazione semantica tentano di imitare il comportamento umano nella creazione di collegamenti ipertestuali che chiariscano il significato dei termini di un documento. Per questo motivo, le principali sfide dell'annotazione semantica sono due: *spotting* e disambiguazione.

1) Lo *spotting* è il procedimento in base al quale vengono scelte le parole, le locuzioni o i sintagmi da annotare tra tutti quelli presenti nel documento. La quantità e la tipologia di espressioni da annotare dipende ovviamente dagli obiettivi del software di annotazione. Possono essere annotati solo i nomi propri (di persone, luoghi, aziende, ecc), in tal caso l'annotazione si dice Named-Entity Recognition, oppure solo le parole di un determinato dominio (per esempio tutti i termini astronomici o tutti i titoli di film). Nel campo del Natural Language Processing ci sono due approcci fondamentali allo *spotting*: quello basato sul Machine Learning e quello basato sulle grammatiche formali. L'approccio basato sul Machine Learning (detto anche statistico), prevede l'annotazione manuale di un certo numero di documenti (*training set*) e l'individuazione di una regola di ap-

prendimento che consenta all'annotatore di svolgere automaticamente il suo compito sulla base degli esempi fornitigli. Il secondo approccio invece fa uso di una grammatica formale (il più delle volte un dizionario, un thesaurus o un'ontologia) che contiene la descrizione delle locuzioni da "spottare" nel testo. L'approccio *dictionary-based* è molto diffuso ed è implementato per esempio in Java dalle librerie LingPipe[94] e in Phyton dal Natural Language Toolkit (NLTK)[95]. Tutte le parole o le locuzioni interessanti per lo *spotting* sono comprese in un dizionario, dunque il software è in grado di estrarre solo elementi compresi in un insieme finito. Ciò è molto utile per delimitare il campo dell'annotazione ad un settore specifico, utilizzando per esempio dizionari specialistici di dominio, o eliminando termini indesiderati perché troppo comuni o ambigui.

2) La disambiguazione (in inglese WSD, Word Sense Disambiguation) è l'operazione di assegnazione di un significato univoco a una parola o un'espressione che può avere diversi sensi a seconda del contesto. Esistono principalmente due tipi di approccio alla WSD (McCarthy, 2009; Navigli, 2009): l'approccio *knowledge-based* e quello supervisionato, o *corpus-based*. Il primo modello sfrutta l'informazione contenuta in risorse lessicali costruite dall'uomo, come dizionari, thesauri o ontologie. Se la risorsa ha una copertura abbastanza ampia e generica, la tecnica sarà applicabile alla maggior parte delle parole esistenti in una lingua, sebbene possa male adattarsi a testi di dominio molto specifico. Basi di conoscenza molto specifiche, di contro, raramente possono assicurare un'ampia copertura di concetti. Le tecniche *knowledge-based* sfruttano le relazioni semantiche (iponimia, iperonimia, sussunzione, ecc.) contenute nel thesaurus per costruire a priori un contesto per un determinato termine o concetto. Il contesto concettuale viene poi confrontato col contesto locale del documento target per disambiguare i termini in esso contenuti. La riuscita di questo approccio è funzione della vastità e della completezza della KB: tradizionalmente, per la lingua inglese, viene utilizzato WordNet, il più grande database lessicale prodotto in forma *machine-processable*. Anche alcuni Linked Open Data *repositories* possono servire da KB per la WSD: in questo caso la presenza di label in RDFS o in SKOS è cruciale, in quanto costituisce l'unico aggancio linguistico della

[94] URL: http://alias-i.com/lingpipe/

[95] URL: http://nltk.org/

KB al testo. L'approccio supervisionato invece prevede l'utilizzo di corpus linguistici annotati a mano per allenare il sistema ad associare direttamente specifici contesti al senso di una parola. I sistemi supervisionati determinano in anticipo quanto contesto deve essere considerato e quali tipi di informazioni del contesto utilizzare. Tali tipi di informazioni includono: 1) le *open-class words* che occorrono a una specifica distanza ai lati della parola target; 2) la *Part-of-Speech* (POS) di queste *open-class words* e della parola target; 3) informazioni sull'argomento del documento target (un documento di medicina, di sport, ecc). Per l'apprendimento vengono usati metodi supervisionati di tipo statistico come il modello Naive Bayes. La letteratura tecnica dimostra che l'approccio supervisionato è ad oggi la tecnica di WSD di maggiore successo (McCarthy, 2009), anche se presenta notevoli problemi di fattibilità. Il tagging manuale di un corpus è costoso, e costruire un sistema che riesce a prendere in considerazione una vasta varietà di contesti per ogni diverso senso di una parola richiede molto tempo.

Con l'avanzare delle tecnologie di Natural Language Processing e di Information Extraction, nell'ultimo decennio si è assistito alla nascita di numerosi progetti di annotazione semantica automatizzata di documenti testuali (Gabrilovich et al., 2006; Kudelka et al., 2007; Mihalcea et al., 2007; Marrero et al., 2010; Mendes at al., 2011; Muñoz-García et al., 2011) . Alcuni di essi sono rimasti circoscritti nel mondo della ricerca, altri hanno trovato uno sbocco sul mercato, incontrando le esigenze dei *Web content publisher* e delle aziende interessate all'*advertising* su Internet. Più recentemente, le tecnologie e la filosofia del Web Semantico hanno preso piede in un crescente numero di settori di interesse per il pubblico, dall'Amministrazione dello Stato all'e-learning, grazie allo sviluppo della comunità dei Linked Open Data e alle già citate iniziative politiche Data.gov e Data.gov.uk. I Linked Data possono essere utilizzati dai sistemi di Information Extraction per fornire basi di conoscenza semantiche, interconnesse e strutturate che possono aumentare la precisione e la *recall*[96] dei meccanismi di annotazione (Rusu et al., 2007; Huang et al., 2009; Muñoz-García et al, 2011; Mendes et al.,

[96] Nell'Information Retrieval la precisione è il rapporto tra i risultati pertinenti ottenuti e i risultati totali, mentre la *recall* è il rapporto tra i risultati pertinenti ottenuti e il totale dei documenti pertinenti esistenti.

2012). Una delle potenzialità maggiori dell'uso di knowledge base come DBpedia si osserva in fase di WSD. Il significato di una parola dipende dal suo contesto e uno strumento come DBpedia fornisce un ricco contesto sia in termini concettuali (le relazioni tra i concetti presenti all'interno dell'ontologia) sia linguistici (morfo-sintattici), grazie all'aggancio diretto con un'enciclopedia tradizionale, ovvero Wikipedia.

2.3.2 Vantaggi nell'uso di DBpedia per il Natural Language Processing

DBpedia, per il fatto di essere una sterminata KB organizzata secondo un modello omogeneo di classi e proprietà e collegata alla più grande fonte di testo enciclopedico multilingue disponibile, rappresenta uno strumento particolarmente interessante per il Natural Language Processing (Giuliano et al., 2009; Ferragina et al., 2010; Ji et al., 2011; Ratinov et al., 2011; Han et al., 2011; Mendes et al., 2012). Di grande utilità sono soprattutto i dataset rilasciati dal progetto DBpedia Spotlight (descritti nel paragrafo 2.2): DBpedia Lexicalizations Dataset, DBpedia Topic Signatures, DBpedia Thematic Concepts e DBpedia People's Grammatical Genders.

Il DBpedia Lexicalization Dataset può essere sfruttato come input di un *dictionary-based spotter* per estrarre le occorrenze delle risorse DBpedia nei testi. Il dizionario viene costruito raccogliendo tutte le *surface form* associate ad ogni risorsa DBpedia, ovvero tutte le varianti linguistiche che esprimono lo stesso concetto in Wikipedia. In questo modo il sistema è in grado di riconoscere locuzioni meno "ufficiali", ma frequenti nel linguaggio comune, come per esempio "la città eterna" per Roma o "il Cavaliere" per Silvio Berlusconi. Questa tecnica è utilizzata sia da DBpedia Spotlight sia da TellMeFirst attraverso l'implementazione di LingPipe (chiamata Exact Dictionary Chunker[97]).

I dataset DBpedia Topic Signatures e DBpedia Thematic Concepts sono invece particolarmente utili per la WSD. Il contesto semantico in cui è inserito il termine da annotare può essere confrontato con le *bag of words* dei candidati per

[97] URL: http://alias-i.com/lingpipe/docs/api/com/aliasi/dict/ExactDictionaryChunker.html

la disambiguazione, in maniera da scegliere il candidato con maggiore similarità concettuale. Secondo l'ipotesi distributiva in linguistica, infatti, le parole che si verificano negli stessi contesti tendono ad avere significati simili (Harris, 1954). L'idea di fondo è che «a word is characterized by the company it keeps» (Firth, 1957), dunque se la parola "Apple" si riferisce all'etichetta fondata dai Beatles è probabile che sia inserita in un contesto dove prevalgano concetti in ambito musicale, mentre se si riferisce all'azienda fondata da Steve Jobs è più probabile che sia associata a concetti nell'ambito dell'elettronica e dell'informatica. Questa strategia per la disambiguazione è stata utilizzata recentemente (Muñoz-García et al., 2011) per l'annotazione basata su DBpedia dei post generati dagli utenti dei social media.

DBpedia People's Grammatical Genders può essere di supporto a un particolare sotto-task dell'annotazione semantica che prende il nome di *coreference resolution*. In linguistica computazionale la *coreference resolution* si occupa di trovare all'interno di un testo le menzioni di una stessa entità sotto forma di elementi grammaticali come pronomi personali, aggettivi e pronomi possessivi, ecc. In una frase di esempio "Petrarch was an Italian writer and poet: his best-known work is the Canzoniere" l'aggettivo "his" deve essere associato correttamente a Francesco Petrarca. Il dataset *People's Grammatical Genders* contiene il genere di ogni entità presente in DBpedia, quindi facilita l'individuazione dei riferimenti alle entità sulla base del loro genere.

2.3.3 Stato dell'arte: una comparazione tra i software di annotazione semantica[98]

I software presentati di seguito sono quelli che hanno superato la "selezione" del mercato, riuscendo a oltrepassare la fase di analisi e di ricerca e proponendosi come soluzioni concrete a determinati problemi dell'utente. Alcuni di essi sono compresi nel test di valutazione condotto dagli autori del progetto DBpedia Spotlight (Mendes at al., 2011) a cui si fa spesso riferimento in questo paragrafo.

[98] Questo studio è stato condotto dal tesista nel corso del progetto TellMeFirst tra il novembre del 2011 e il febbraio del 2012. Le informazioni contenute si riferiscono alla versione dei software (e della loro documentazione) disponibile in tale periodo.

Gli approcci sono alquanto diversificati, tuttavia hanno in comune l'utilizzo di una KB, più o meno compatibile con gli standard del Semantic Web, e più o meno proprietaria, per l'estrazione dei termini e la loro disambiguazione. Wikify! (Mihalcea, 2007), pur non avendo raggiunto lo stadio di progetto concreto, è stato inserito nella comparazione per il suo indubbio interesse metodologico.

Per condurre la loro valutazione, gli autori di Spotlight hanno costruito un corpus di news preannotato a partire da 30 articoli selezionati casualmente dal sito Web del New York Times[99] appartenenti a 10 diverse categorie di argomenti. Il *gold standard* utilizzato per la comparazione, disponibile online per il download[100], è prodotto dagli annotatori stessi che partecipano al test, come lista degli URI sui quali c'è maggiore accordo tra gli annotatori. Si legge nella documentazione di Spotlight:

> In order to construct a gold standard, each evaluator first independently annotated the corpus, after which they met and agreed upon the ground truth evaluation choices. The ratio of annotated to not-annotated tokens was 33%. The annotation agreements are significantly different from those expected by chance. After this, the four annotators met and decided on a single set of annotations to keep, generating an entity set shared in the file: gold.set.[101]

Nel seguito vengono esaminati i principali software di annotazione presenti sul mercato, prendendo in considerazione per ognuno di essi l'utilizzabilità, le prestazioni e la documentazione. Il capitolo 3 della tesi illustra la soluzione proposta da TellMeFirst, le sue basi metodologiche e il motivo delle sue scelte architetturali.

Nome	**Precisione**	**Recall**	**F1 score[102]**	**Demo online**	**Open Source**	**Italiano**	**Punti di forza**

[99] URL: http://www.nytimes.com/

[100] URL: http://spotlight.dbpedia.org/download/isemantics2011-evaluation.zip

[101] URI: http://wiki.dbpedia.org/spotlight/isemantics2011/evaluation

[102] In statistica l'F1 score è la media armonica tra precisione e recall di un test.

Alchemy API	0,5	0,1	14,7%	sì	no	no	-
Open Calais	0,4	0,1	16,7%	sì	no	no	Ottimo il riconoscimento dell'argomento del testo, nonostante il link ai LOD non sia sempre presente (KB proprietaria).
Zemanta	0,8	0,3	39,1%	sì	no	no	Enhancement dei contenuti sull'interfaccia intuitivo e rapido.
Machine Linking FBK	0,5	0,7	59,5%	sì	no	sì	Prestazioni di precisione e recall molto alte, italiano presente ma migliorabile.
Wikify!	-	-	-	no	no	-	Approccio metodologicamente interessante alla disambiguazione: copresenza di tecniche knowledge-based e corpus-based.
Apache Stanbol	-	-	-	sì	sì	no	Modularità, possibilità di attivare o disattivare motori di annotazione e disambiguazione diversi. Lo stack tecnologico di language detection e di document parsing si basa su librerie open source.
DBpedia Spotlight	0,8	0,6	56,0%	sì	sì	no	Configurabilità, ottimo utilizzo dei LOD (DBpedia), chiarezza nella do-

							cumentazione e nel codice. Risultati di precisione e recall per la lingua inglese molto buoni, con possibilità di customizzare il sistema per l'italiano.

Tabella 1 – Quadro riassuntivo del confronto tra annotatori

2.3.3.1 AlchemyAPI[103]

E' un prodotto nato nel 2008 all'interno dell'azienda americana Orchestr8, che ha base a Denver (Colorado). Si tratta di un software che utilizza tecniche di NLP e di Machine Learning per analizzare contenuti testuali ed estrarre metadati semantici, ovvero informazioni su luoghi, persone, aziende e argomenti presenti nel testo. AlchemyAPI possiede un meccanismo di Named Entity Disambiguation basato su tecniche statistiche e su di un vasto dataset proprietario, contenente persone, luoghi, aziende, ecc. Questo dataset conserva anche informazioni contestuali utili per la disambiguazione (ad esempio, per le aziende, la collocazione, il nome dei dirigenti più importanti, i principali prodotti, ecc). Si legge nella documentazione:

> A series of statistical algorithms are combined with a huge data-set describing the world's objects, individuals, and locations.[...] Our disambiguation engine employs tens of millions of hints describing traits of the world's objects, individuals, and locations. We employ a variety of public and non-public data-sets. Hints vary depending on the specific type of entity being disambiguated. For example, when disambiguating people, we utilize information on a person's career, where they're located, who they work for, and so on. For companies: key executives, notable products, industry, location, etc.[104]

I Linked Data vengono sfruttati in fase di arricchimento, ovvero per collegare le entità estratte ad ulteriori contenuti provenienti da DBpedia, Freebase, US

[103] URL: http://www.alchemyapi.com/

[104] URL: http://www.alchemyapi.com/api/entity/disamb.html

Census, GeoNames, UMBEL, OpenCyc, YAGO, MusicBrainz, CIA Factbook and CrunchBase.

AlchemyAPI è un software proprietario, espone delle Web service API e una demo online. Fornisce varie soluzioni business e una soluzione free accessibile via API. Non è chiaro se la demo online abbia le stesse caratteristiche della soluzione business. Sta di fatto che l'italiano non è supportato, per quanto nella documentazione si dica che il prodotto supporta «English, French, German, Italian, Portuguese, Russian, Spanish, and Swedish»[105]. L'utilizzo del Web service è gratuito solo se non si superano le 1.000 richieste al giorno (30.000 per gli enti di ricerca). AlchemyAPI fornisce spesso un link a DBpedia, tra le altre risorse utilizzate in fase di arricchimento, e per questo è stato utilizzato senza ulteriori modifiche nel test di valutazione di DBpedia Spotlight[106]. C'è da obiettare che per AlchemyAPI il link a DBpedia è solo un risultato accessorio, perché, come Open Calais, il sistema dispone di un dataset proprietario. Quindi la comparazione effettuata dagli autori di Spotlight non rende del tutto merito della qualità dell'annotazione di questo prodotto. La recall infatti risulta molto bassa, mentre la precisione è nella media. F1 score: 14,7. Precision: 0,5. Recall: 0,1.

2.3.3.2 Open Calais[107]

Open Calais è un progetto nato nel 2008 in seno al colosso americano Thomson Reuters[108]. Si presenta come un Web Service che arricchisce automaticamente contenuti testuali con metadati semantici, usando tecniche di NLP e di Machine Learning. Open Calais utilizza, per l'estrazione delle *named entities*, una propria KB proprietaria, il cui schema concettuale è stato rilasciato pubblicamente come ontologia OWL[109]. La KB di Open Calais contiene un gran numero di entità (soprattutto persone, organizzazioni, prodotti commerciali, luoghi ed eventi) e rispetta gli standard dei Linked Data, pur non essendo accessibile via

[105] URL: http://www.alchemyapi.com/api/entity/langs.html

[106] Qui e nel seguito i risultati del test di valutazione di Spotlight si intendono tratti da Mendes et al. (2011).

[107] URL: http://www.opencalais.com/

[108] URL: http://thomsonreuters.com/

[109] URL: http://www.opencalais.com/documentation/opencalais-web-service-api/calais-ontology-owl

SPARQL endpoint. Gli URI al suo interno sono dereferenziabili e alcuni di essi sono linkati ai nodi di DBpedia, Freebase e Geonames. Ecco un esempio per la città di Parigi in Francia: http://d.opencalais.com/er/geo/city/ralg-geo1/797c999a-d455-520d-e5cf-04ca7fb255c1.html.

Per alcune categorie (Aziende, Prodotti ed Entità Geografiche), Open Calais utilizza un meccanismo automatico di disambiguazione basato sul contesto. Si legge nella documentazione:

> OpenCalais now supports three types of entity disambiguation: Company disambiguation, Geographical disambiguation and Product (Electronics) disambiguation. […]
>
> Company disambiguation uses a proprietary Thomson Reuters database of companies as its reference set and is tuned for public companies. […] Geographical disambiguation uses Freebase as its reference set. […] Disambiguation of electronic products currently uses the Shopping.com catalogue as its reference set.[110]

La disambiguazione è spiegata purtroppo in termini molto generici:

> To resolve an entity, we will use the entity name itself, and possibly additional contextual clues appearing in the text around it. For example, when resolving a company name, clues in the text as to what the company does or where it is located may help to resolve among several possible identities of the company.[111]

Anche Open Calais è un software proprietario, espone web service API e una demo online. L'italiano non è supportato (l'interfaccia ritorna il messaggio: «Calais continues to expand its list of supported languages, but does not yet support your submitted content»). Le API sono a pagamento se si vogliono processare più di 50.000 documenti al giorno. La piattaforma non fornisce sempre un collegamento tra le entità della propria KB proprietaria e gli URI di DBpedia. Gli autori di DBpedia Spotlight dunque, per permettere la valutazione di Open Calais, hanno adoperato un semplice algoritmo di trasformazione delle risorse proprietarie in entità di DBpedia basato sulle label di Open Calais (una risorsa di Open Calais con label "Apple" diventa "dbpedia:Apple"). Appare abbastanza

[110] URL: http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/entity-disambiguation

[111] Ibidem.

chiaro che questo meccanismo di trasformazione delle entità estratte in entità di DBpedia finisce col risultare svantaggioso per Open Calais, che utilizza una KB proprietaria. Questo è uno dei motivi per cui, nel test di valutazione di Spotlight il software è indicato con indici di recall e precisione piuttosto bassi. F1 score: 16,7. Precision: 0,4. Recall: 0,1.

2.3.3.3 Zemanta[112]

Zemanta è un'azienda *venture-backed* di New York fondata nel 2007. Il prodotto omonimo è un software in grado di annotare semanticamente pagine Web e linkarle con contenuti esterni come immagini, testi e video. Questi contenuti provengono da Wikipedia, IMDB, MusicBrainz e Amazon Books. Zemanta è stato pensato in particolar modo per i blogger e gli editori di contenuti digitali, per questo è stato rilasciato per prima cosa come plug-in di Firefox e successivamente come plug-in dei principali CMS, come Drupal, Wordpress e Joomla. Successivamente il software è stato esposto tramite Web service e sono state fornite apposite API per gli sviluppatori. Nella (carente) documentazione di Zemanta si legge che il processo di disambiguazione è eseguito per mezzo del contesto delle parole, ma non vengono forniti ulteriori dettagli. Si legge:

> Understanding of text is used to find out whether specific phrase should be linked to specific page. In one context Zemanta may decide a connection is desirable in the other the same link might not be suggested. Context is also used to disambiguate between different possibilities when linking (for example Apollo the space program and Apollo the Greek god).[113]

Si tratta di un software proprietario che espone delle Web service API e una demo online. La lingua italiana non è supportata (i termini italiani sono ricondotti agli omografi inglesi). Le API di Zemanta non ritornano esplicitamente un URI di DBpedia, ma un link alla pagina di Wikipedia da cui l'URI può essere facilmente inferito. Nel test di valutazione di Spotlight, Zemanta domina in precision, ma ha una recall bassa. F1 score: 39,1. Precision: 0,8. Recall: 0,3.

---

[112] URL: http://www.zemanta.com/

[113] URL: http://developer.zemanta.com/media/files/docs/zemanta_api_companion.pdf

2.3.3.4 Machine Linking FBK[114]

Machine Linking FBK (in precedenza nota come "Wiki Machine") è un progetto italiano di Claudio Giuliano e Paolo Lombardi, nato all'interno della Fondazione Bruno Kessler[115] e diventato una start-up alla fine del 2010. La tecnologia di Machine Linking analizza i testi generati dagli utenti del Web e si rivolge in particolare ai gestori di servizi di Social Networking che devono organizzare *user-generated content* in lingue diverse. Il software permette di associare dei tag semantici alle parole chiave e ai nomi propri all'interno di un testo. È multilingue e funziona anche per testi brevi fino a 100 caratteri. La Fondazione Bruno Kessler opera come partner tecnologico dell'azienda, con i suoi 50 ricercatori e dottorandi dell'unità di Comprensione del Linguaggio Naturale.

Machine Linking FBK è un software proprietario, non espone un Web service, ma solo una demo online. La lingua italiana è supportata. La demo non ritorna esplicitamente URI di DBpedia, ma collegamenti alle pagine di Wikipedia da cui gli URI possono essere facilmente inferiti. Il sistema ha ottenuto un punteggio molto alto nel test di valutazione di DBpedia Spotlight, ma ha la tendenza a privilegiare la *recall* a scapito della precisione. F1 score: 59,5%. Precision: 0,5. Recall: 0,7. La documentazione è pressoché assente[116].

2.3.3.5 Wikify!

Progetto di ricerca nato nel 2007 all'interno dell'Università del North Texas, Wikify! ha come principali autori Rada Mihalcea (affermato ricercatore nell'ambito della WSD) e Andras Csomai. Si tratta di un sistema software che simula l'annotazione manuale degli articoli all'interno di Wikipedia. Sulla base del *Manual of Style*[117] di Wikipedia, che specifica le regole da rispettare per linkare un articolo ad altre risorse dell'enciclopedia, gli autori di Wikify! hanno progettato le funzionalità del proprio software in modo che il risultato finale fosse il

[114] URL: http://thewikimachine.fbk.eu/
[115] URL: http://www.fbk.eu/it
[116] URL: Le informazioni sono state prese da:
http://www.smau.it/milano11/partners/machinelinking/
http://thewikimachine.fbk.eu/html/index.html
[117] URL: http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

più possibile indistinguibile da un'annotazione umana *wiki-style*. Questo significa che l'attenzione degli sviluppatori si è concentrata maggiormente sulla precisione piuttosto che sulla *recall*. Nel manuale di stile infatti è esplicitamente detto che solo i concetti utili a fornire una conoscenza più approfondita sull'argomento della voce devono essere interlinkati e che quelli poco pertinenti devono essere tralasciati. Si legge nella documentazione di Wikify!:

> Although prepared for human annotators, these guidelines [il manuale di stile di Wikipedia] represent a good starting point for the requirements of an automated system, and consequently we use them to design the link identification module for the Wikify! system. The main recommendations from the Wikipedia style manual are highlighted below: (1) Authors/annotators should provide links to articles that provide a deeper understanding of the topic or particular terms, such as technical terms, names, places etc. (2) Terms unrelated to the main topic and terms that have no article explaining them should not be linked. (3) Special care has to be taken in selecting the proper amount of keywords in an article – as too many links obstruct the readers' ability to follow the article by drawing attention away from important links. (Mihalcea, 2007, p. 3)

Vi è dunque una fase di raffinamento delle *features* del documento già durante lo spotting, prima della disambiguazione. Wikify! in sostanza non estrae tutte le entità in fase di spotting, ma solo quelle che soddisfano determinati criteri di rilevanza all'interno del documento. Lo spotting è eseguito, come in DBpedia Spotlight, a partire da un dizionario esatto di termini costituito dai titoli degli articoli di Wikipedia e dai *wikilinks*. Ma già in fase di spotting a ogni termine estratto viene assegnato un ranking che ne indica la rilevanza all'interno dell'articolo secondo i seguenti 3 criteri:

1) il rapporto tra il numero di occorrenze del termine in quel dato documento e il numero di documenti in cui il termine appare deve superare una certa soglia;
2) il termine o l'espressione deve comparire nel documento più spesso di quanto non comparirebbe per caso.
3) il rapporto tra il numero di documenti in cui il termine è stato già selezionato come parola chiave e il numero totale di documenti in cui il termine è apparso deve superare una certa soglia.

A livello di WSD, Wikify! prevede l'implementazione e l'integrazione di due soluzioni diverse.

1) La prima, di tipo *knowledge-driven*, si basa sul confronto del contesto della forma di superficie estratta (il paragrafo del testo in cui compare) con il contesto che questa entità ha nella knowledge base, ovvero la corrispondente pagina di Wikipedia. Si legge:

> For instance, given the context "it is danced in 3/4 time, with the couple turning 180 degrees every bar", and assuming that "bar" could have the meanings of bar music or bar counter, we process the Wikipedia pages for the music and counter meanings, and consequently determine the sense that maximizes the overlap with the given context. (Mihalcea, 2007, p. 6)

2) La seconda, di tipo *corpus-driven*, prevede l'utilizzo di tecniche di apprendimento supervisionato e di un classificatore di tipo Naive Bayes. Si crea un *training feature vector* per ogni occorrenza della parola da disambiguare in una pagina di Wikipedia (sotto forma di *wikilink*). Questo vettore è composto da elementi che riguardano sia il contesto locale (le pagine di Wikipedia dove la *surface form* appare come *wikilink*) sia quello globale o *topical* (la pagina di Wikipedia dedicata all'entità che rappresenta il senso della *surface form*). Di seguito una schematizzazione del vettore:

la surface form	il part-of-speech tag della surface form	i part part-of-speech tag delle parole vicine alla surface form	un contesto locale di tre parole a sinistra e a destra della surface form	un contesto globale costituito da massimo 5 parole che occorrono almeno 3 volte nella pagina di Wikipedia dedicata all'entità/senso

Tabella 2 - Training feature vector in Wikify!

Ogni *wikilink* presente in Wikipedia è associato a un vettore di questo tipo. L'insieme dei vettori viene utilizzato per allenare un classificatore

Naive Bayes. Ogni elemento da disambiguare viene rappresentato con un vettore analogo per consentire il confronto col *training set*.

Wikify! non espone una demo online né un Web service, ma è ben documentato. Sebbene non sia stato inserito nel confronto tra gli annotatori di DBpedia Spotlight, proprio per la mancanza di API pubbliche, gli autori di Spotlight fanno notare come il numero delle entità annotato da Wikify! sia molto basso per volontà stessa dei progettisti del software (le ragioni sono state esposte in precedenza). Le entità "sopravvissute" allo *spotting* sono relativamente poche e semplici da disambiguare perché sono le più attinenti al contesto complessivo del documento:

> Wikify!'s conservative spotting only annotate 6% of all tokens in the input text. In Wikify!, this spotting yields surface forms with low ambiguity for which even a random disambiguator achieves an F1 score of 0.6. (Mendes, 2011, p. 7).

2.3.3.6 Apache Stanbol[118]

Stanbol è un progetto austriaco nato nel marzo 2010 col nome di FISE (Furtwangen IKS Semantic Engine) all'interno della comunità IKS (Interactive Knowledge Stack)[119], cofinanziata dall'Unione Europea e dal Salzburg Research Institute[120]. Attualmente è in incubazione presso l'Apache Incubator. Si tratta di un framework open source che ha l'obiettivo di facilitare l'implementazione di Content Management System semantici attraverso appositi moduli Java esposti come API REST. Stanbol eccede gli obiettivi di un software di annotazione semantica, puntando su una serie di funzionalità a più ampio raggio:

1) Persistenza: un'insieme di servizi che archiviano informazioni semantiche e le rendono ricercabili.
2) *Enhancement*: servizi che aggiungono nuova informazione ai contenuti testuali.
3) *Reasoning*: per controllare la consistenza dell'informazione archiviata o per arricchirla tramite inferenza.

[118] URL: http://stanbol.apache.org/
[119] URL: http://www.iks-project.eu/
[120] URL: http://www.salzburgresearch.at/

4) Interazione con l'utente: sviluppo di interfacce grafiche intelligenti per fruire dei servizi semantici.

A livello di *document annotation*, Stanbol fornisce strumenti per coprire l'intero processo di annotazione, appoggiandosi ad altre librerie open source della Apache Foundation. L'identificazione della lingua viene fatta per mezzo della libreria Apache Tika[121]; l'estrazione del plain text dal documento è eseguita con Aperture[122]; le operazioni di NLP, ovvero *sentence detection*, *POS tagging*, *tokenization* e *chunking*, vengono effettuate per mezzo di Apache OpenNLP[123]; OpenNLP è utilizzato anche per la Named Entity Recognition (NER), secondo il modello della Massima Entropia[124], per riconoscere luoghi, persone e organizzazioni.

La NER di OpenNLP è affiancata da un meccanismo di spotting dictionary-based. Stanbol consente un certo grado di configurabilità a livello della knowledge base da adottare, in quanto prevede l'utilizzo di un proprio "EntityHub" come interfaccia ai dataset presenti in locale o in remoto. Di default, l'EntityHub di Stanbol è costituito da una copia parziale locale di DBpedia. Si legge nella documentazione:

> The Entityhub provides two main services. The Entityhub provides the connection to external linked open data sites as well as using indexes of them locally. Its services allow to manage a network of sites to consume entity information and to manage entities locally. A small index of approx. 43k entities from DBpedia comes with the default installation.[125]

Le label dell'EntityHub nella lingua riconosciuta dal Language Identification Engine vengono utilizzate come dizionario per lo spotting. Nel matching tra i sintagmi, alcuni elementi come le preposizioni vengono scartati, per aumentare la recall.

---

[121] URL: http://tika.apache.org/
[122] URL: http://aperture.sourceforge.net/
[123] URL: http://incubator.apache.org/opennlp/
[124] URL: http://en.wikipedia.org/wiki/Maximum_entropy_method
[125] URL: http://incubator.apache.org/stanbol/docs/trunk/entityhub.html

> All tokens (of the text) following the current position are searched within the label. As of now, tokens MUST appear in the correct order within a label (e.g. "Murdoch Rupert" will NOT match "Rupert Murdoch").On the first processable token of the text that is not present within the label matching is canceled. […] On the second non-processable token not found in the label the matching is also canceled (e.g. "University of Michigan" will match "University Michigan").[126]

Il modulo che si occupa dello spotting dictionary-based è il Keyword Linking Engine. Esso gestisce anche i *redirect*, ovvero le voci del dizionario che reindirizzano ad altre voci, lasciando all'utente la libertà di configurare il meccanismo in base alle sue esigenze. L'utente può scegliere di eliminare il reindirizzamento, di seguirlo o di elencare semplicemente le opzioni disponibili:

> In some cases suggested entities might redirect to others. In the case of Wikipedia/DBpedia this is often used to link from acronyms like IMF to the real entity International Monetary Fund. But also some Thesauri define labels as own Entities with an URI and users might want to use the URI of the Concept rather than one of the label. To support such use cases the KeywordLinkingEngine has support for redirects. Users can first configure the redirect mode (ignore, copy values, follow) and secondly the field used to search for redirects (default=rdfs:seeAlso). If the redirect mode != ignore for each suggestion the Entities referenced by the configured redirect field are retrieved. In case of the "copy values" mode the values of the name, and type field are copied. In case of the "follow" mode the suggested entity is replaced with the first redirected entity.[127]

Lo stesso EntityHub viene utilizzato da un modulo chiamato Named Entity Tagging Engine per collegare i termini estratti alle risorse di DBpedia o di altri dataset della LOD, come GeoNames. Nel caso una di queste risorse sia ambigua, il risultato finale, ovvero il documento annotato, indica la lista di tutti i candidati affiancati da un ranking che ne rappresenta la *confidence*.

> For the sentence "John Smith lives in London", you will get several EntityAnnotations for the terms "London", "John Smith" form your linking target resource (in this

[126] URL:
http://incubator.apache.org/stanbol/docs/trunk/enhancer/engines/keywordlinkingengine.html
[127] Ibidem.

case DBpedia) together with a confidence value, which can be used to sort the suggestions.[128]

Da uno scambio di mail tra gli sviluppatori di Stanbol, si evince che questo *ranking* è calcolato *a priori* dalla quantità di link a quella risorsa presenti in DBpedia, e dunque non ha nessun collegamento con il contesto.

> Even if you plan to index all entities you might want to use entity scores, because such scores are also used to boost entities within the Entityhub. So if you search, than results will be sorted by the number of incoming links within DBpedia. This ensures that a search for "Paris" returns Paris France as best result. Without such boosts Paris Texas could be also returned as best result.[129]

Insieme al Named Entity Tagging Engine, altri moduli vengono forniti per l'enhancement. Questi moduli sfruttano le API di servizi esterni, come Zemanta e Open Calais, per annotare semanticamente il testo. Possono essere attivati parallelamente al modulo di default (il Named Entity Tagging Engine) o separatamente, perché i risultati vengono integrati tutti nella stessa risposta.

Il framework è completamente open source ed espone un Subversion online[130]. Il codice compila e deploya correttamente, risultando non particolarmente difficile da installare. Stanbol mette a disposizione online ben tre istanze di demo, con un numero diverso di moduli attivati, per consentire all'utente di provare il software con diverse configurazioni. I linguaggi per adesso supportati non includono l'italiano, ma sono comunque numerosi: inglese, tedesco, danese, svedese, olandese e portoghese. Il motivo è da individuare nella fase di *POS tagging*. OpenNLP infatti richiede un training del proprio *POS tagger* con un corpus pre-annotato nella lingua specifica (*treebank*). In italiano un corpus di questo genere non è reperibile gratuitamente online.

---

[128] URL: http://incubator.apache.org/stanbol/docs/trunk/enhancer/engines/namedentitytaggingengine.html

[129] URL: http://mail-archives.apache.org/mod_mbox/incubator-stanbol-dev/201111.mbox/%3CF158714B-FEC9-403B-AD65-7920FB11264A@gmail.com%3E

[130] URL: http://svn.apache.org/repos/asf/incubator/stanbol/trunk/

Apache Stanbol non è incluso nel test di valutazione di DBpedia Spotlight, né viene citato dagli autori di Spotlight nella sezione sullo stato dell'arte. Pur essendo nato nel 2010 e potendo contare su una comunità di 30 sviluppatori, Stanbol ha una documentazione piuttosto carente e male organizzata. Il progetto trova uno dei maggiori *supporter* nell'azienda Nuxeo[131], specializzata nella produzione di sistemi di Enterprise Content Management. Nuxeo è stata anche la prima azienda ad integrare Stanbol all'interno di un prodotto commerciale, la Nuxeo Enterprise Platform. I risultati delle demo dipendono dal tipo di configurazione: le prime due, che utilizzano solo i moduli più stabili sembrano avere una precisione e una *recall* piuttosto modeste per la lingua inglese, non al livello degli altri competitor. La più recente installazione[132], che attiva anche i moduli sperimentali, ha una buona *recall*, ma pecca in precisione. Il punto debole di Stanbol sembra proprio essere, sia leggendo la documentazione sia provando le demo, la WSD.

2.3.3.7 DBpedia Spotlight

Spotlight è un progetto di ricerca nato nel giugno del 2010 all'interno della Freie Universität Berlin[133]. Il gruppo di sviluppatori è costituito da Pablo Mendes, Max Jakob e Jo Daiber ed è supervisionato da Chris Bizer, il fondatore del progetto DBpedia, nonché figura di grande rilievo a livello internazionale nel campo del Semantic Web. Spotlight è finanziato dalla Commissione Europea attraverso il progetto LOD2 – Creating Knowledge out of Linked Data[134].

Spotlight si propone come uno strumento per annotare automaticamente le menzioni di entità di DBpedia all'interno di testi e quindi per collegare dati non strutturati ai repository della Linked Open Data Cloud (Mendes et al., 2011; Mendes et al., 2012). Pone grande rilevo al problema della WSD, cercando di offrire all'utente una soluzione costumizzabile che incontri le sue esigenze a livello sia di precisione che di recall.

---

[131] URL: http://www.nuxeo.com/en
[132] URL: http://dev.iks-project.eu:8081/
[133] URL: http://www.fu-berlin.de/en/index.html
[134] URL: http://lod2.eu/Welcome.html

L'architettura di Spotlight è modulare: si tratta di un progetto Maven sviluppato in Java+Scala e scomponibile in 4 sottoprogetti buildabili autonomamente: Indexing, Core, Evaluation e RESTful API. Il modulo di Indexing si occupa della costruzione dei dataset necessari al sistema per compiere le operazioni di annotazione semantica presenti nel modulo Core. Il modulo di Evaluation è un framework che consente di effettuare alcuni test di valutazione delle performance di Spotlight e di altri annotatori accessibili via Web service. Il modulo RESTful contiene le API del Web service di Spotlight.

Il modulo di Indexing prende in input:

1. DBpedia Labels: le etichette in una specifica lingua di tutte le risorse presenti in DBpedia;
2. DBpedia Redirects: gli URI di DBpedia che reindirizzano verso altri URI;
3. DBpedia Disambiguation: gli URI di DBpedia che rappresentano pagine di disambiguazione di Wikipedia, ovvero indicano i possibili significati di una stessa entry.
4. DBpedia Ontology Infobox Types: il valore dell'attributo rdf:type di ogni URI presente in DBpedia
5. Wikipedia XML Dump: il *dump* in XML di tutti i contenuti di Wikipedia in una specifica lingua.

È importante notare che i dataset estratti da DBpedia sono specifici per ogni lingua, cioè contengono URI di DBpedia differenti per le diverse versioni di DBpedia. Questo potrebbe apparire contrario alla logica di una KB multilingua, ove a medesimi URI sono associate varie etichette linguistiche. Il team di DBpedia ha scelto questa soluzione per consentire di utilizzare la conoscenza specifica delle versioni di Wikipedia (inglese, francese, italiana, ecc.) attraverso KB specifiche per le diverse lingue. Tuttavia questo tipo di scelta può influire in maniera negativa sulle prestazioni di Spotlight, in particolar modo se si utilizza una versione di DBpedia poco estesa come quella italiana (DBpedia in italiano contiene un numero di URI otto volte inferiore rispetto a DBpedia in inglese).

Il modulo di Indexing produce tre importanti output intermedi per il progetto:

1) ConceptURIs.list: una lista di tutti gli URI di DB che identificano concetti, quindi pulita dalle risorse di disambiguazione e di reindirizzamento.
2) Redirects_tc.tsv: un mapping tra gli URI di reindirizzamento e gli URI concettuali in DBpedia.
3) Surface_forms-Wikipedia-TitRedDis.tsv: tutte le surface form con accanto le risorse di DBpedia candidate a costituirne il senso.

La prima colonna del file Surface_forms-Wikipedia-TitRedDis.tsv viene utilizzato come dizionario per lo *spotting*: Spotlight utilizza come *spotter* un LingPipe Exact Dictionary-Based Chunker, basato sull'algoritmo di string maching Aho-Corasick[135].

Lo *spotting* ritorna una lista di oggetti di tipo SurfaceFormOccurrence.scala, ognuna dei quali ha come attributi una SurfaceForm, un Context (il paragrafo da dove la SurfaceForm è stata estratta) e un TextOffset (un intero che descrive la posizione della *surface form* nel testo).

La disambiguazione segue un procedimento abbastanza complesso. Prima di tutto viene processato il *dump* di Wikipedia per estrarre tutte le occorrenze dei *wikilink* insieme ai relativi contesti. Queste informazioni vengono inizialmente conservate in una lista di oggetti di tipo DBpediaResourceOccurrence.scala, ognuno dei quali rappresenta un *wikilink* ed ha i seguenti attributi:

1) id: l'identificativo univoco di una singola occorrenza del *wikilink*;
2) resource: L'URI della risorsa target di DBpedia;
3) support: un intero che indica il numero di volte in cui quel *wikilink* occorre in Wikipedia;
4) types: la/e classe/i a cui appartiene la risorsa target di DBpedia;
5) surface form: la surface form del *wikilink*;
6) context: una stringa che contiene il paragrafo del *wikilink*;
7) textOffset: un intero che indica la posizione della *surface form* nel testo;

La lista di DBpediaResourceOccurrence è poi trasferita su un file TSV, dove ogni *wikilink* si trova su una diversa riga e le righe sono ordinate per URI. A partire da questo CSV (occs.csv) si costruisce un indice Lucene[136], a cui vengono ag-

[135] URL: http://en.wikipedia.org/wiki/Aho%E2%80%93Corasick_string_matching_algorithm
[136] URL: http://lucene.apache.org/

giunte in fase di reindicizzazione le *surface form* contenute in Surface_forms-Wikipedia-TitRedDis.tsv, affinché una risorsa sia la candidata di tutte le possibili *surface form* che la rappresentano. L'indice Lucene viene poi compattato ed utilizzato per la disambiguazione, comparando gli attributi dei SurfaceFormOccurrence con i campi dell'indice (scaricabile anche indipendentemente dal progetto[137]).

Il risultato del processo di annotazione è una lista di oggetti di tipo DBpediaResourceOccurrence.scala, che hanno come attributi surfaceForm le forme di superficie estratte dal testo e come attributi Resource le risorse di DBpedia disambiguate.

Il codice di Spotlight è open source ed esposto online tramite Subversion[138]: l'installazione non presenta particolari difficoltà. C'è una demo per la versione 0.5[139], inoltre i servizi sono raggiungibili via API REST. L'unica lingua per ora supportata è l'inglese, anche se ci sono gruppi di sviluppatori che stanno lavorando alla customizzazione del software per lo spagnolo, il portoghese e il tedesco[140].

La configurabilità è uno dei vantaggi principali di DBpedia Spotlight rispetto agli altri software di annotazione semantica. Calibrando il *trade-off* tra precisione e *recall*, l'utente può ottenere dal sistema i risultati più adatti alle proprie necessità. Dal test di valutazione condotto dagli sviluppatori stessi del progetto, il software è risultato molto competitivo, posizionandosi al secondo posto dopo Machine Linking FBK. Nel diagramma di comparazione (vedi Figura 1) la performance di Spotlight non è rappresentata da un punto, ma da una linea in funzione dei parametri di *confidence* (sicurezza nella disambiguazione, a sua volta funzione del numero di candidati e della similarità contestuale) e *support* (numero di wikilinks che rimandano alla specifica risorsa di Wikipedia). Valori alti di *confidence* e *support* fanno arrivare Spotlight ad una precisione dello 0.8, mentre valori bassi di *confidence* e *support* lo conducono ad una *recall* dello 0.6. Nella sua configurazione ottimale, dunque, il sistema raggiunge un punteggio F1 di 56.0%.

[137] URL: http://spotlight.dbpedia.org/download/release-0.4/full-index.tgz
[138] URL: http://dbp-spotlight.svn.sourceforge.net/viewvc/dbp-spotlight/
[139] URL: http://spotlight.dbpedia.org/demo/index.html
[140] URL: http://wiki.dbpedia.org/spotlight/internationalization/

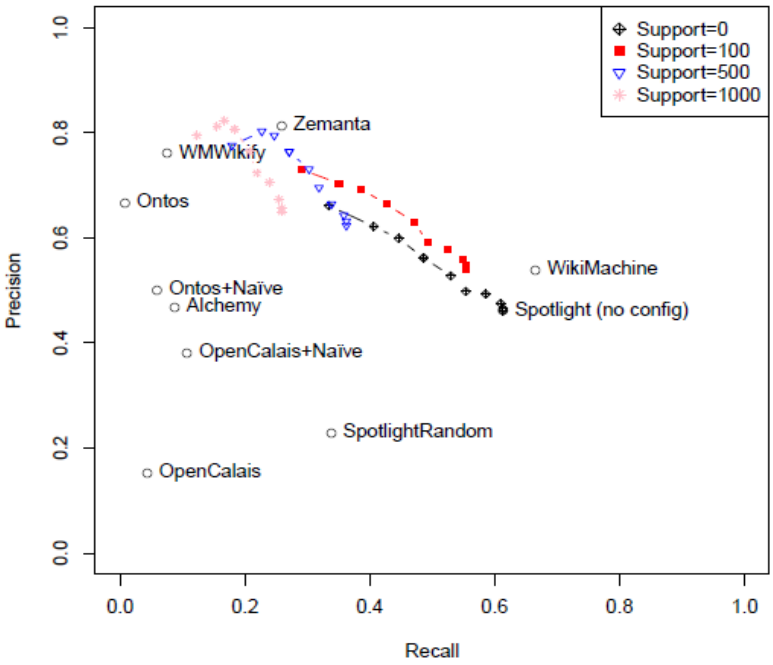


Figura 10 - Diagramma di comparazione di DBpedia Spotlight (Mendes, 2011, p. 6)

Capitolo 3

TellMeFirst: un sistema per l'annotazione, la classificazione e l'arricchimento dei documenti attraverso DBpedia

3.1 L'approccio di TellMeFirst all'annotazione e alla classificazione semantica

Il progetto TellMeFirst[141] (TMF) ha avuto inizio nel mese di ottobre 2011 grazie ad un finanziamento "Working Capital - Premio Nazionale Innovazione" da parte di Telecom Italia. A partire dal luglio del 2012 è stato sviluppato prevalentemente all'interno del Centro Nexa su Internet & Società[142] del Dipartimento di Informatica e Automatica del Politecnico di Torino.

TellMeFirst è un software per la classificazione e l'arricchimento automatico di documenti. L'annotazione semantica è solo un obiettivo secondario del sistema, ma rappresenta una delle funzionalità più complesse e sfidanti in termini di tecnologie da gestire e di *know-how* prodotto. Il focus di TMF è sui documenti testuali (HTML, PDF, Word, ecc.) in lingua italiana e inglese. Particolare attenzione è rivolta alla WSD, senza la quale i tentativi di classificazione del testo in base al suo argomento risulterebbero inefficaci.

Come altri software dello stesso genere (DBpedia Spotlight, Apache Stanbol, visti in precedenza), TMF sceglie di utilizzare DBpedia come KB di riferimento

[141] URL: http://tellmefirst.polito.it/
[142] URL: http://nexa.polito.it/

per l'estrazione e la disambiguazione dei contenuti. Lo strumento DBpedia è stato scelto per la classificazione in quanto corpus di Wikipedia è perfetto come *training set* per ogni approccio basato sul Machine Learning. Dal punto di vista dell'annotazione, DBpedia si presta come strumento adatto a supportare il tipo di approccio *knowledge-based* per i seguenti motivi:

1) È una vasta base di conoscenza, con un ottimo livello sia di copertura sia di specificità. Inoltre è aggiornata costantemente con le entità e i fatti più recenti.
2) È ricca di relazioni semantiche tra i concetti, compresa la loro classificazione in categorie sia generiche (Person, Organization, Place, ecc.) sia specifiche (French Poet, Fruit Tree, ecc.).
3) È provvista di label multilingua che consentono di agganciare la KB ai documenti in italiano e in inglese.

Anche l'approccio supervisionato all'annotazione semantica può beneficiare notevolmente da DBpedia, grazie al collegamento diretto tra DBpedia e il vasto corpus pre-annotato multilingua costituito da Wikipedia. Le annotazioni sono in questo caso i *wikilinks*, ovvero i link interni a Wikipedia che collegano le diverse pagine tra loro. Per questo motivo, avendo a disposizione una vasta risorsa linguistica e semantica come DBpedia/Wikipedia, TellMeFirst può implementare al proprio interno i due diversi tipi di approccio, e utilizzarli separatamente o in parallelo, a seconda della necessità.

Il disambiguatore di TMF ha, di fatto, tre sotto-componenti: un Knowledge-based Disambiguator, un Corpus-based Disambiguator e un First Sense Heuristic Disambiguator. Quest'ultimo è pensato per entrare in funzione in quei casi in cui i primi non abbiano fornito un risultato accompagnato da un livello sufficiente di *confidence*. Si tratta di un meccanismo che sfrutta il "coefficiente di *prominence*" delle risorse, ovvero il numero di volte in cui vengono citate in Wikipedia attraverso un wikilink, per decidere sul significato più comune di un termine ambiguo. Quando un termine non è disambiguato dal Knowledge-based Disambiguator e dal Corpus-based Disambiguator con un certo grado di *confidence*, allora il First Sense Heuristic Disambiguator gli assegna il significato più comune. Per quanto semplicistico, l'approccio euristico è stato dimostrato essere spesso solo pochi

punti percentuali sotto i sistemi di WSD con più elevate prestazioni (McCarthy, 2009).

Dal punto di vista della classificazione del testo, l'approccio scelto per TellMeFirst è di tipo *memory-based learning*, una sottocategoria del *lazy learning*. Caratteristica distintiva di tale approccio, detto anche *instance-based*, è che il sistema non si occupa di creare un modello astratto delle categorie di classificazione (profili) prima del processo di categorizzazione del testo, ma assegna il documento target a una classe sulla base di un confronto locale tra i documenti preclassificati e quello target (Cheng et al., 2009). Questo significa che il classificatore deve avere in memoria tutte le istanze del *training set* e calcolare in fase di classificazione la distanza vettoriale tra i documenti di training e il documento non classificato. Tale approccio viene inserito nella famiglia del *lazy learning* (letteralmente "apprendimento pigro"), in opposizione all'*eager learning* (letteralmente "apprendimento impaziente"): mentre il primo rimanda alla fase di classificazione (*consultation time*) il calcolo della similarità col *training set*, il secondo anticipa questa operazione alla fase di apprendimento (*training time*), dove appunto vengono creati i profili specifici delle categorie e decisa la funzione in base alla quale effettuare la classificazione (Sammut et al., 2010, p. 571).

Nello specifico, l'algoritmo utilizzato da TellMeFirst è il k-Nearest Neighbor (kNN), un tipo di approccio *memory-based* che sceglie la categoria o le categorie di appartenenza del documento target sulla base dei k documenti più simili a quello target nello spazio vettoriale (Sammut et al., 2010, p.714). La variabile k è sempre uguale a 1 in TellMeFirst, quindi la classe scelta è quella del documento con maggiore similarità rispetto al documento target.

Il training set è costituito da tutti i paragrafi in cui occorre un *wikilink* all'interno di Wikipedia. Questi paragrafi vengono memorizzati in un indice Lucene, come Field CONTEXT di documenti che rappresentano risorse di DBpedia. Nell'indice a ogni risorsa DBpedia (dunque a ogni pagina Wikipedia) corrisponde un Lucene Document, e in ogni Document vi sono tanti Field CONTEXT quanti sono i paragrafi in cui appare un *wikilink*. Al momento della classificazione (seguendo l'approccio *lazy*) il documento target viene trasformato in una query Lucene booleana sul campo CONTEXT dell'indice per scoprirne la somiglianza con-

cettuale coi contesti delle voci di Wikipedia. Per il calcolo della similarità, viene utilizzata la Default Similarity[143] di Lucene, che combina il modello booleano al modello vettoriale (Vector Space Model, SVM) dell'Information Retrieval: i risultati approvati dalla ricerca booleana sull'indice, vengono poi ordinati in base a SVM. Lucene si occupa dello *stemming*, della lemmatizzazione e del filtro (attraverso apposite *stop words* per l'italiano e l'inglese) delle *feature* sia dei documenti di *training* sia del documento target trasformato in query. La query e i documenti di *training* diventano entrambi vettori pesati di *feature* (secondo il modello della *bag of words*), dove il peso di ogni *feature* viene calcolato in base all'algoritmo TF-IDF (vedi paragafo 2.3.1). La query ritorna una lista di documenti (ovvero di URI DBpedia) ordinata secondo il *similarity score*. La formula di *scoring* è la seguente:

$$\text{cosine-similarity}(q,d) = \frac{V(q) \cdot V(d)}{|V(q)|\,|V(d)|}$$

Dove q è la query, d il documento di training, V(q) il vettore pesato della query e V(d) il vettore pesato del documento di *training*. Al numeratore vi è dunque il prodotto scalare tra il vettore della query e il vettore del documento di *training*, mentre al denominatore la distanza euclidea tra il vettore della query e il vettore del documento di *training*[144]. Ottenuta la lista ordinata di risultati, si applica il metodo RCut per il *thresholding* (Yang, 2001), tenendo solo i primi 7 risultati in base al loro *rank* e scartando gli altri.

La tecnica utilizzata da TellMeFirst per la classificazione, basandosi sul modello dello spazio vettoriale per la raffigurazione sia dei documenti di *training* che del documento target, può essere considerata di tipo "spaziale" (*spatial technique*). La similarità tra due documenti può essere vista geometricamente come la distanza tra i due vettori che rappresentano i documenti in uno spazio vettoriale *n*-dimensionale dove *n* è il numero di *feature* dell'intero corpus di *training* (Figura 11). Il modello dello spazio vettoriale è anche alla base delle librerie Lucene.

[143] URL: http://lucene.apache.org/core/3_6_0/api/all/org/apache/lucene/search/Similarity.html

[144] Ibidem.

Lucene è ottimizzata per svolgere in tempi rapidissimi il calcolo della distanza tra i documenti secondo l'algoritmo TF-IDF: data un query che rappresenta le *feature* del documento target, in pochi istanti è ritornata una lista di documenti simili indicizzati, anche quando l'indice è popolato da milioni di documenti. Lo *score* ottenuto con Lucene rappresenta l'inverso della distanza tra due documenti: più alto è lo *score*, più vicini sono i documenti nello spazio vettoriale.

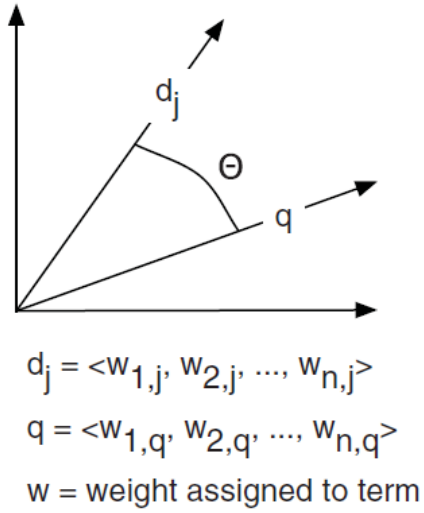


Figura 11 - Rappresentazione della distanza tra due documenti in uno spazio vettoriale bidimensionale (Ingersoll, 2013, p. 49)

3.2 Componenti del sistema

TMF ha un'architettura modulare per consentire il riuso di ogni componente e/o l'inserimento di nuovi moduli software (anche di terze parti) che approfondiscano o estendano le sue funzionalità. Nello schema riportato sotto, sono indicati i componenti del sistema come passi fondamentali del processo dall'acquisizione all'arricchimento del documento in input.

Il componente **C1** è un *document parser* che utilizza librerie come Apache PDFBox[145], Apache POI[146], Snacktory[147] per l'estrazione per estrarre il *plain text* da documenti in vari formati, PDF, Word, HTML, ecc .

[145] URL: http://pdfbox.apache.org/
[146] URL: http://poi.apache.org/

C2 è un wrapper che utilizza uno o più software/servizi esterni (per esempio Freeling[148]), indicati con **C3**, per eseguire il *POS tagging* del testo e la lemmatizzazione dei nomi comuni e degli aggettivi. Questo procedimento è particolarmente importante per evitare errori dovuti all'omografia di nomi comuni, aggettivi e nomi propri (fermi / [Enrico] Fermi, leopardi / [Giacomo] Leopardi, ecc.)

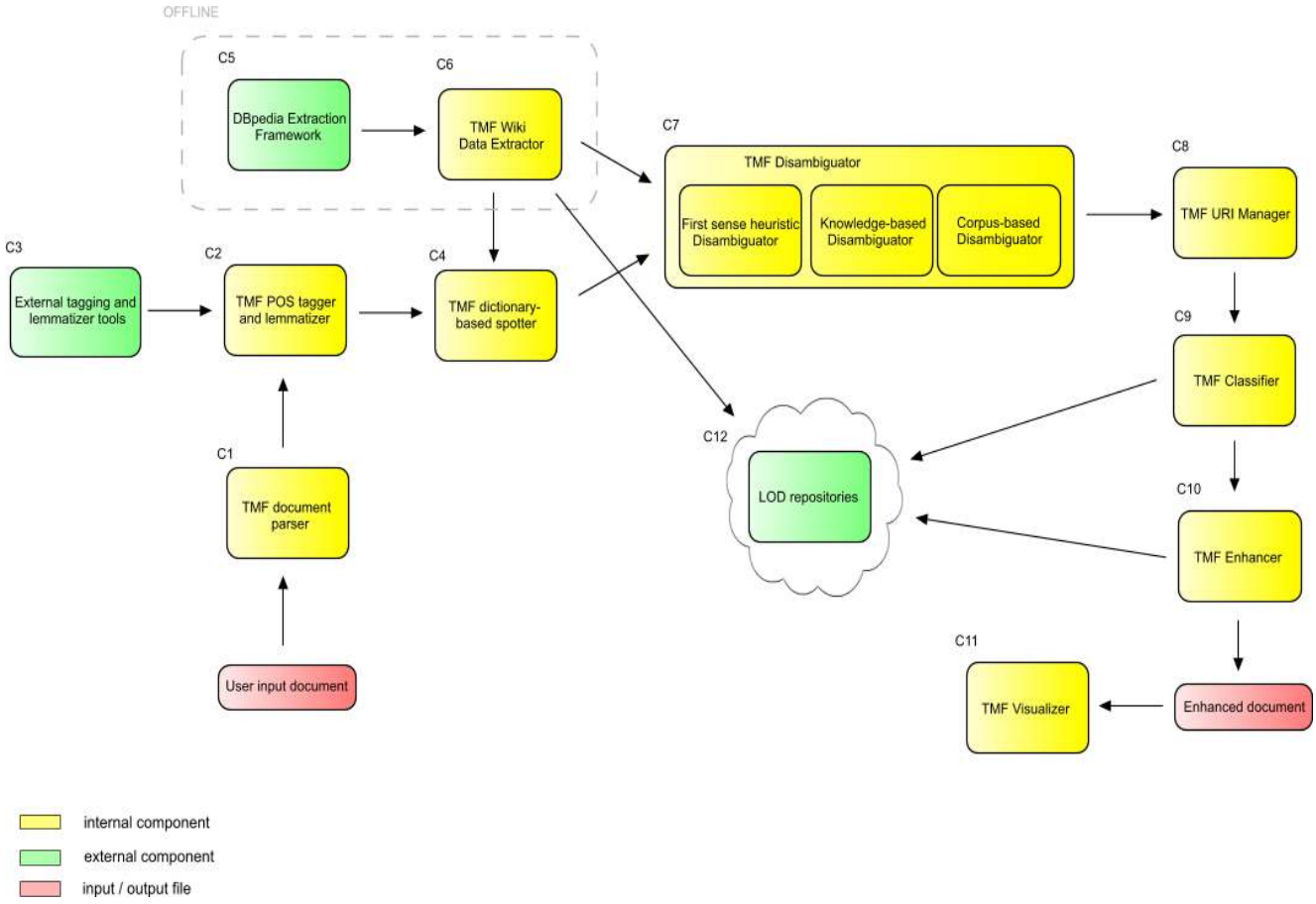


Figura 12 - Architettura semplificata di TellMeFirst

C2 è un wrapper che utilizza uno o più software/servizi esterni (per esempio Freeling[149]), indicati con **C3**, per eseguire il *POS tagging* del testo e la lemmatizzazione dei nomi comuni e degli aggettivi. Questo procedimento è particolarmen-

[147] URL: https://github.com/karussell/snacktory
[148] URL: http://nlp.lsi.upc.edu/freeling/
[149] URL: http://nlp.lsi.upc.edu/freeling/

te importante per evitare errori dovuti all'omografia di nomi comuni, aggettivi e nomi propri (fermi / [Enrico] Fermi, leopardi / [Giacomo] Leopardi, ecc.)

C4 è il componente che si occupa dello *spotting*, ovvero dell'estrazione dal testo dei termini da annotare. Fa uso di un dizionario interno (popolato dal componente C6) per estrarre solo i termini che matchano con le voci del dizionario stesso. Un'implementazione del *dictionary-based spotter* è fornita per esempio dalle librerie Java LingPipe.

C6 è il modulo che si occupa dell'elaborazione dei dati provenienti da DBpedia e da Wikipedia, necessari a TMF per lo *spotting* e la disambiguazione. C6 utilizza la componente esterna **C5**, ovvero il DBpedia Extraction Framework, per ricavare alcuni dataset per la lingua italiana non disponibili online. Entrambi i moduli funzionano offline come *preprocessing*, contrariamente a tutti gli altri moduli che operano *runtime*.

C7 è il componente di disambiguazione, di cui si parla più approfonditamente nel paragrafo 3.4.

C8 si occupa del mapping tra le risorse della versione inglese di DBpedia (quella accessibile attraverso lo SPARQL endpoint pubblico[150]) e le risorse corrispondenti della versione italiana di DBpedia (i cui dataset sono scaricabili dal sito del progetto[151]). In uscita dal componente C7, le risorse appena disambiguate sono nella forma "http://it.dbpedia.org/resource/NomeRisorsa", in quanto TellMeFirst utilizza i dataset italiani per la disambiguazione, ma in uscita da C8 sono nella forma "http://dbpedia.org/resource/ResourceName" e quindi utilizzabili per le successive query sui Linked Open Data.

C9, il modulo di classificazione, utilizza i LOD repositories **C12** (per adesso DBpedia, ma non è escluso l'utilizzo di altri Linked Data in fasi successive del progetto) per stabilire quale sia (o quali siano) l'argomento (o gli argomenti) del testo. A partire da una serie di concetti di DBpedia (la lista delle entità in output dal Disambiguator) C9 lancia una serie di query SPARQL sull'endpoint di DBpedia per trovare la (o le) entità che hanno un numero maggiore di collegamenti (object properties) con i concetti estratti dal documento target.

[150] URL: http://dbpedia.org/sparql

[151] URL: http://it.dbpedia.org/

Il componente di enhancement, **C10**, a partire dalle entità di DBpedia che sono state individuate come argomento del testo, fa ancora uso dei Linked Open Data (e di altri servizi esterni) per ricavare nuove informazioni e nuovi contenuti da aggiungere al documento per arricchirlo.

C11 è il modulo di visualizzazione del sistema e si occupa di raccogliere in una unica interfaccia, in maniera appropriata secondo il tipo di contenuto, le nuove informazioni con cui il documento di partenza è stato arricchito e che sono utili per comprendere meglio il significato del testo.

3.3 Interazione tra componenti e artefatti

Di seguito è illustrata l'architettura completa del sistema, ovvero l'interazione tra i componenti del sistema e gli artefatti intermedi da esso prodotti.

3.3.1 Dataset iniziali

A1 e **A2** sono i dump in formato XML degli interi corpora di Wikipedia in italiano e in inglese. Essi sono scaricabili direttamente dalla pagina di download di Wikipedia[152].

Anche gli artefatti A3, A4 e A5 e A16 sono disponibili online, in quanto fanno parte dei risultati del progetto DBpedia[153]. **A3** è la lista degli URI di DBpedia che rimandano a pagine di disambiguazione; **A4** la lista dei wikilinks compresi in ogni pagina di Wikipedia; **A5** la lista di tutte le label delle risorse della versione italiana di DBpedia; **A16** di quella inglese[154].

[152] URL: http://dumps.wikimedia.org

[153] URL: http://wiki.dbpedia.org/Downloads37

[154] Per gli artefatti in cui non è specificata la lingua tra parentesi (es. A3, A4, ecc.) è da intendersi che viene utilizzata la versione di DBpedia inglese per l'inglese, la versione italiana per l'italiano. Utilizzare i dataset della versione italiana di DBpedia è risultato conveniente per poter sfruttare in fase di spotting i redirect e le pagine di disambiguazione specifiche per la lingua italiana.

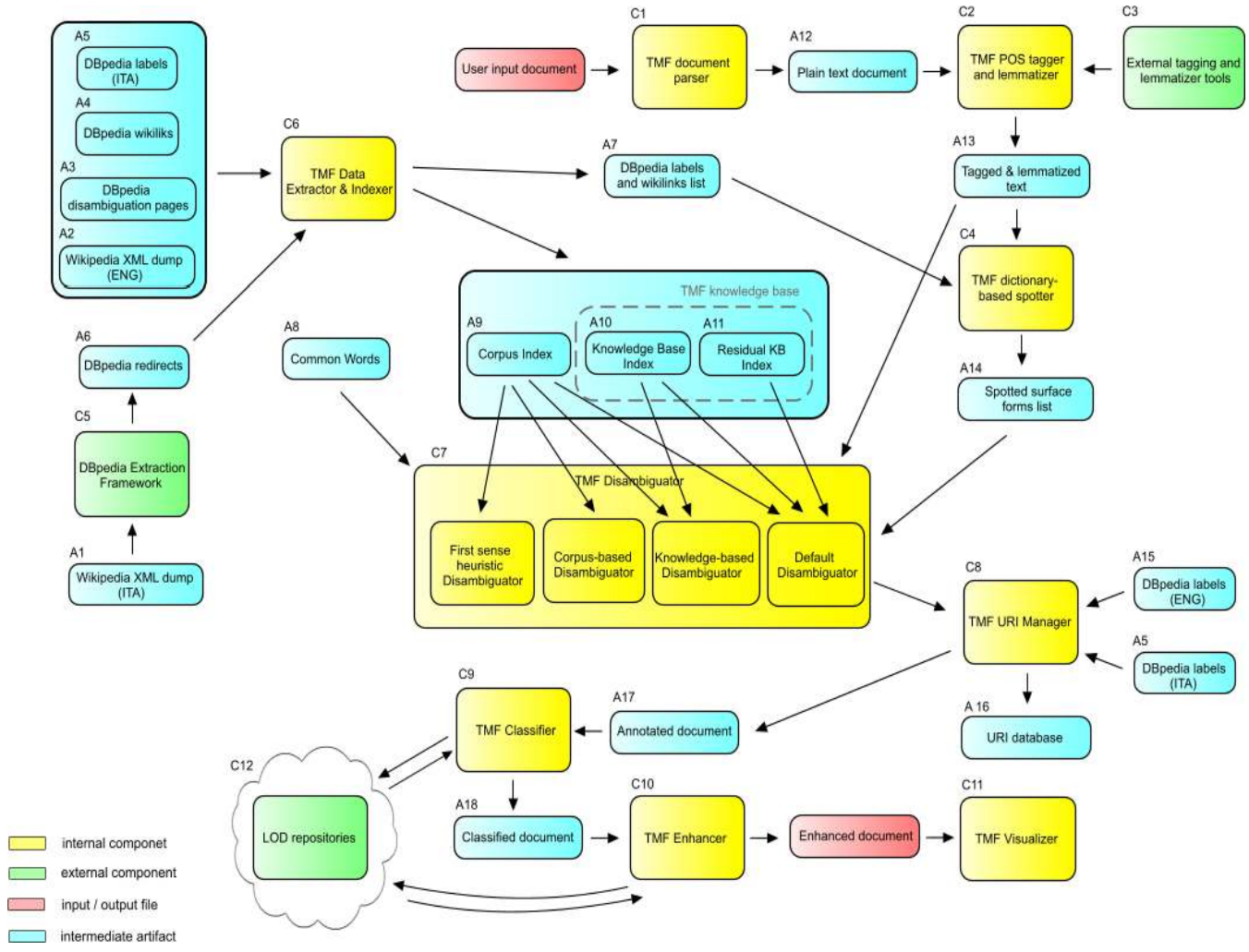


Figura 13 - Architettura completa di TellMefirst

A6, la lista di tutte gli URI di DBpedia che non costituiscono un'entità a sé, ma rimandano ad URI di altre entità, è disponibile online per la lingua inglese[155], ma non per quella italiana. Per questo motivo è necessario creare il dataset italiano per mezzo del DBpedia Extraction Framework (C5), dando in input al componente il *dump* in italiano di Wikipedia (A1).

[155] Ibidem.

A8 è una lista delle parole più comuni nella lingua scelta. Per l'italiano è stato utilizzato l'elenco dei lemmi del corpus LIP, uno dei principali risultati del progetto BADIP (BAnca Dati dell'Italiano Parlato[156]). Questo artefatto viene utilizzato in fase di annotazione per scartare a priori i termini più ricorrenti della lingua scelta e quindi meno interessanti per comprendere il contesto concettuale del documento target. Tale procedimento è molto comodo non solo per evitare di annotare le parole di uso più comune in una determinata lingua, e quindi meno interessanti dal punto di vista concettuale, ma anche per escludere un set arbitrario di parole a scelta del progettista (ad esempio parole che possono essere considerate molto ambigue in un determinato dominio). Il servizio REST di annotazione di TellMeFirst prevede un apposito parametro da passare nella richiesta per attivare o disattivare questa funzionalità.

A1 (o A2), A3, A4, A5 e A6 finiscono in input alla componente TMF Wiki Data Extractor (C6). A8 va in input al disambiguatore (C7), mentre A5 e A16 vengono utilizzate dall'URI Manager (C8) per costruire il database A17.

3.3.2 Dataset elaborati

Il componente C6 ritorna come output i dataset necessari al funzionamento dei moduli di spotting (C4) e di disambiguazione (C7). Si tratta dei 6 artefatti sotto elencati. I dettagli dell'utilizzo degli artefatti A9, A10 e A11 da parte dei disambiguatori di TellMeFirst sono specificati nel paragrafo 3.

DBpedia labels and wikilinks list (**A7**): un LingPipe MapDictionary[157] contenente la lista di tutte le label di DBpedia in una certa lingua, più tutte le surface form dei wikilink estratti dalla versione di Wikipedia nella stessa lingua. Tale dizionario è compilato dalle librerie LingPipe in maniera da poter essere utilizzato da un LingPiepe ExactDictionaryChunker[158]. A7 viene infatti utilizzato dal TMF Dictionary-Based Spotter (C4), che appunto utilizza ExactDictionaryChunker come classe di base.

[156] URL: http://badip.uni-graz.at/index.php
[157] URL: http://alias-i.com/lingpipe/docs/api/com/aliasi/dict/MapDictionary.html
[158] URL: http://alias-i.com/lingpipe/docs/api/com/aliasi/dict/ExactDictionaryChunker.html

Corpus Index (**A9**): è l'indice Lucene di riferimento per il Corpus-based Disambiguator, ma viene utilizzato anche dagli altri disambiguatori per filtrare i candidati secondo criteri morfosintattici. Si veda fig. 3. Il Corpus Index associa ad ogni URI di DBpedia i seguenti campi:

1) SURFACE FORMS: le label e i redirect di quella risorsa, a cui possono essere aggiunte anche le forme di superficie dei suoi wikilink.
2) PREFERRED SURFACE FORM: la parola o espressione che occorre più volte come forma di superficie dei wikilink di quella risorsa.
3) CONTEXTS: le porzioni di testo in cui è contestualizzato ogni wikilink di quella risorsa.
4) URI COUNT: il numero di wikilink di ogni risorsa, ovvero quante volte si fa riferimento a quella risorsa attraverso un link in Wikipedia.
5) TYPE: il valore del tag rdf:type della risorsa in DBpedia.

Knowledge Base Index (**A10**): è l'indice Lucene di riferimento per il Knowledge-based Disambiguator. Esso associa ad ogni URI di DBpedia un campo "KB", che contiene tutti gli URI che compaiono almeno due volte come *wikilinks* nella pagina Wikipedia di riferimento. A10 è pensato per essere l'insieme delle *bag of concets* di tutte le risorse di DBpedia.

Residual KB Index (**A11**): un indice Lucene che viene utilizzato dal Default Disambiguator per aumentare la *recall* rispetto al Knowledge-base Disambiguator. Associa ad ogni risorsa di DBpedia un campo "KB" contenente tutti gli URI che compaiono una sola volta come *wikilinks* nella pagina Wikipedia di riferimento della risorsa.

Il componente C8 produce invece due artefatti: uno è ottenuto per mezzo di una elaborazione offline (A16), l'altro invece è fornito runtime alla componente C9 (A17). A16 è un database MySql costituito da una sola tabella "urimap" di tre colonne "itaUri", "label" e "engUri". È in sostanza un mapping tra le risorse della versione inglese di DBpedia e le risorse corrispondenti della versione italiana, basato sulla corrispondenza delle label in lingua italiana. Queste label sono

infatti univoche, perché corrispondono ai titoli degli articoli di Wikipedia, i quali sono univoci per scelta editoriale[159].

C8 utilizza il database A16 per ottenere A17, ovvero il documento annotato con le risorse della versione inglese di DBpedia (es: http://dbpedia.org/resource/Carnival) a partire dal documento annotato con le risorse della versione italiana di DBpedia (es: http://it.dbpedia.org/resource/Carnevale). Laddove non esista una corrispondente risorsa inglese della risorsa italiana richiesta, C8 fornisce l'URL dell'articolo di Wikipedia in italiano (es: http://it.wikipedia.org/wiki/Carnevale). A17 viene poi dato in input al componente di classificazione C9.

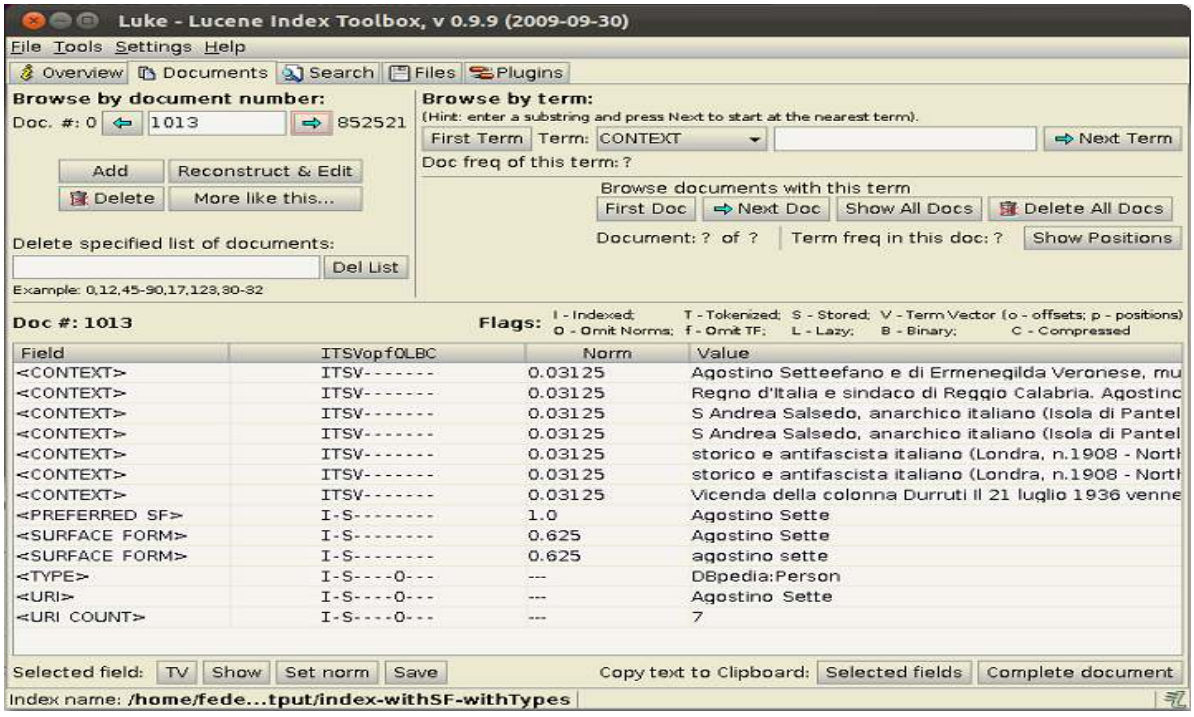


Figura 14 - Corpus Index visualizzato con Luke sulla risorsa "Agostino_Sette"

[159] Si veda il manuale di stile di Wikipedia (http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style) e la pagina specifica sui titoli degli articoli (http://en.wikipedia.org/wiki/Wikipedia:Article_titles).

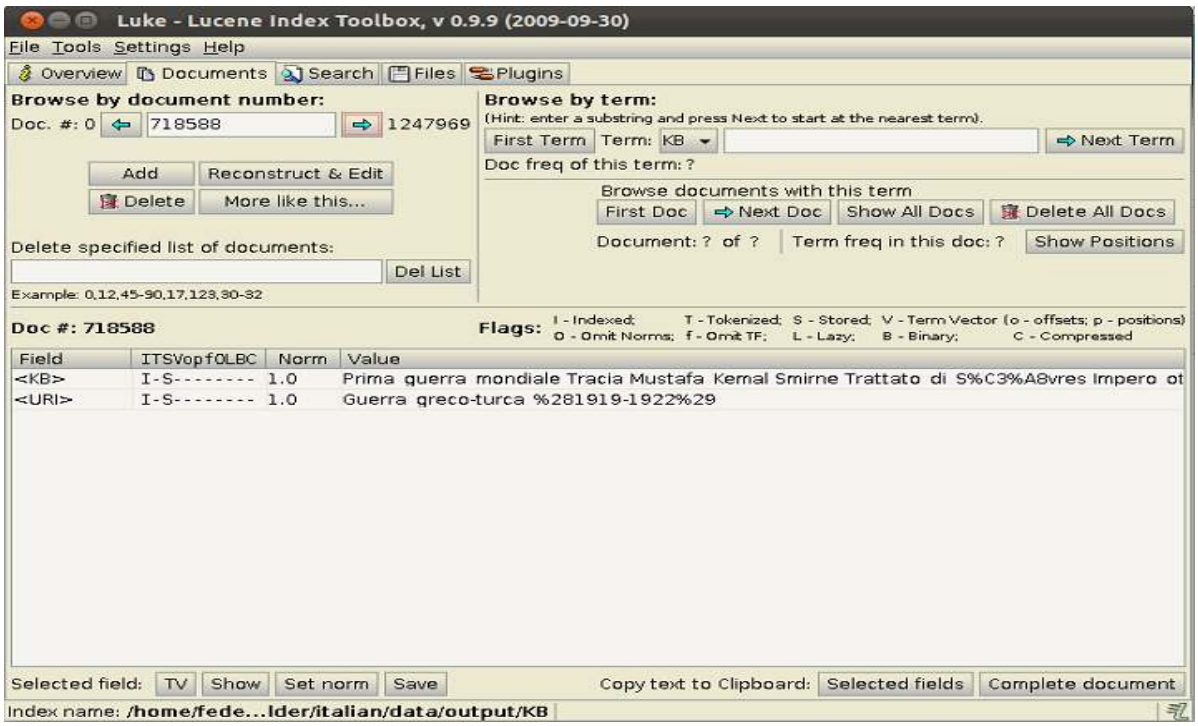


Figura 15 - KB Index visualizzato con Luke sulla risorsa "Guerra_greco-turca_(1919-1922)"

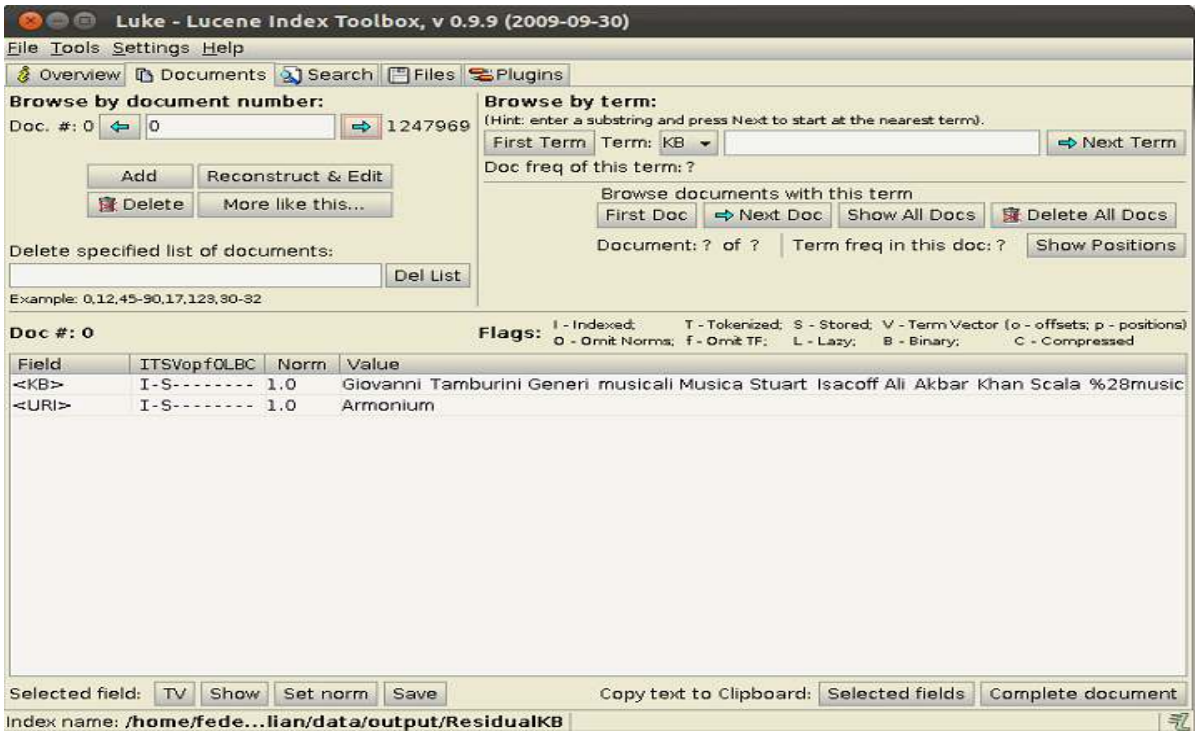


Figura 16 - Residual KB Index visualizzato con Luke sulla risorsa "Armonium"

Il documento in input al sistema viene inizialmente parsificato dal componente C1 di TMF per produrre un file di *plain text* (**A12**). Quest'ultimo passa attraverso il componente C2, che ne estrae la lista di *named entities* e di lemmi per i nomi comuni e gli aggettivi (**A13**). A13 viene utilizzato sia dal componente di *spotting* C4, sia dal componente di disambiguazione C7, in quanto alcuni controlli morfosintattici sui candidati vengono effettuati sulla base dei *POS tag* dei *token* del testo target.

In output dal TMF Dictionary-based Spotter (C4) abbiamo la lista delle *surface form* estratte dal documento target (**A14**), pronte per la disambiguazione.

Gli artefatti A13 e A14, insieme ad A8 e A9, A10 e A11 visti in precedenza, finiscono in input al componente C7 di disambiguazione, il quale produce un documento annotato coi riferimenti a risorse non ambigue di DBpedia (**A17**). Il componente di classificazione C9 di TMF processa A17 e ritorna lo stesso file con l'aggiunta di annotazioni riguardanti il suo argomento complessivo, sempre nella forma di DBpedia URI (**A18**). L'Enhancer C9 annota ulteriormente A18 con le nuove informazioni tratte dai Linked Open Data, informazioni che vengono poi gestite dalla componente di visualizzazione del sistema (C10).

3.4 Il modulo di disambiguazione di TellMeFirst

Il disambiguatore di TMF si compone di quatto diversi moduli, per effettuare il processo di disambiguazione in modo differente a seconda delle esigenze dell'utilizzatore. È l'utente stesso dunque a scegliere il tipo di disambiguazione, inserendo il parametro "disambiguator" nella chiamata REST al servizio di annotazione di TMF, tra quelli disponibili:

1) Corpus-based
2) Knowledge-based
3) First Sense Heuristic
4) Default

La Tabella 3 riassume il risultato dei test degli annotatori di TellMeFirst effettuati su un corpus di 10 stralci di articoli del quotidiano online La Repubbli-

ca[160]. L'ultima colonna indica i possibili scenari di utilizzo dei diversi sistemi di annotazione. In grassetto le performance migliori per ogni colonna.

Disambiguator	Average time per word	Average precision	Average recall	Use scenario
Corpus-based	0,04 s.	0,85	0,21	Annotazione "live" di articoli di giornale o di blog, consentendo magari all'utente di eliminare manualmente le annotazioni ritenute scorrette.
Knowledge-based	0,07 s.	0,99	0,05	Classificazione automaticamente dei documenti di un archivio o di un'attività online in base a tag concettuali connessi a DBpedia, Ontology-based document retrieval.
First sense heuristic	0,04 s.	0,78	0,24	Analogo a quello del Corpus-based Disambiguator, in ambiti meno specifici e più generici, dove ha senso suggerire come primo significato di una parola quello più comune in Wikipedia.
Default	0,10 s.	0,96	0,08	Annotazione offline, estrazione delle principali informazioni dai testi, classificazione automatica ed enhancement dei documenti.

Tabella 3 - Test sui disambiguatori di TMF

I quattro sottocomponenti del modulo di disambiguazione hanno sia logica specifica per la propria categoria, sia logica in comune, condividendo alcuni metodi e finzionalità. Di seguito andiamo ad approfondire il funzionamento e la composizione di ognuno di essi, delineandone anche i possibili scenari d'uso. Le illustrazioni X, 4, 5 e 6 schematizzano graficamente i differenti processi di disambiguazione: i cerchi rappresentano un'azione, mentre i rettangoli un risultato intermedio o finale del processo.

[160] URL: http://www.repubblica.it/

3.4.1 Corpus-based disambiguation

La disambiguazione *corpus-based* sfrutta il Corpus Index di TMF per scegliere i candidati i cui contesti in Wikipedia hanno maggiore aderenza col documento target. Il contesto della *spotted surface form*, ovvero l'intero testo sottoposto a TMF, viene confrontato attraverso una query Lucene con i contesti dei candidati presenti nel Corpus Index, ricavando una *ranked list* dei risultati più attendibili. I principali passi della disambiguazione Corpus-based sono illustrati nella figura 3.

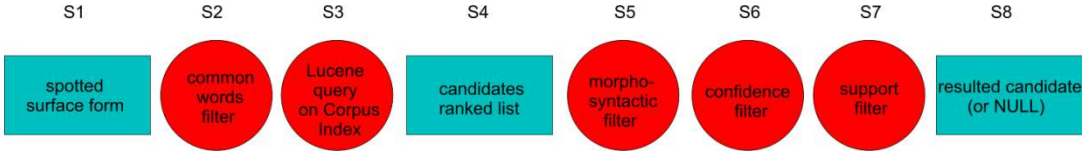


Figura 17 - Processo di annotazione corpus-based

La surface form individuata in fase di *spotting* (**S1**) passa inizialmente per il *common words filter* (**S2**), che elimina a priori le parole contenute nel file Common Words (A8). Se la *surface form* passa indenne dal *common words filter,* viene utilizzata come elemento costitutivo di una query Lucene sul Corpus Index (**S3**). Questa query va a cercare la *surface form* nei campi SURFACE FORMS e PREFERRED SURFACE FORM dell'indice, mentre cerca l'intero testo da annotare nel campo CONTEXTS dell'indice stesso. Il risultato è una *ranked list* dei candidati (**S4**) per quella *surface form*.

La lista dei candidati è sfoltita da tre filtri successivi (*morphosyntactic filter*, *confidence filter*, *support filter*), che eliminano i candidati che non soddisfano i loro requisiti.

Il *morphosyntactic filter* (**S5**) opera dei controlli sul candidato sfruttando congiuntamente il POS tag della *surface form* e i campi PREFERRED SURFACE FORM e TYPE del Corpus Index. I criteri di controllo sono i seguenti:

1) I token che iniziano con la minuscola non devono essere annotati con risorse che hanno la PREFERRED SURFACE FORM maiuscola.

2) I token che iniziano con la minuscola, e che sono composti da una sola parola, non devono essere annotati con risorse che contengano nel campo TYPE le stringhe “Person”, “Place”, "Work", "Music", "Film"o "Single"[161].
3) Se un token ha l'articolo in minuscolo, non deve essere associato a una risorsa con l'articolo maiuscolo nella PREFERRED SURFACE FORM.
4) Se un token non ha l'articolo, non deve essere associato a una risorsa che comincia per un articolo nella PREFERRED SURFACE FORM.

Il *confidence filter* (**C6**) è comandato dall'utente stesso per mezzo della GUI di TMF: si può settare un valore di confidence da 0.0 a 0.9, rendendo il filtro più o meno severo. Il *confidence filter* si basa su un mapping tra i possibili valori di *confidence* e i possibili valori di score dei risultati della query Lucene. Dunque valori di *confidence* bassi elimineranno valori di score bassi in modo proporzionale. Questo filtro controlla anche la distanza percentuale tra il primo e il secondo candidato per la disambiguazione: perciò valori di confidence bassi elimineranno i candidati che hanno una distanza percentuale proporzionalmente bassa dal loro secondo candidato.

Il *support filter* (**C7**) è anch'esso comandato dall'utente nella GUI di TMF. Nella *textfield* del *support* l'utente può scrivere un numero compreso tra 0 e 999, in quanto il filtro si basa sul valore del campo URI COUNT del Corpus Index. La cifra inserita dall'utente viene confrontata con il valore di URI COUNT della risorsa candidata, così se la cifra è più bassa il candidato viene scartato.

In uscita da questi filtri, dunque, vi è una nuova lista di candidati (**C8**). Questa può essere di lunghezza 0, e quindi la *surface form* di partenza non sarà annotata, oppure di lunghezza 1, e in questo caso l'unico candidato rimasto sarà il risultato dell'annotazione, o ancora può essere di lunghezza >1. In tal caso, il primo candidato della lista sarà il risultato dell'annotazione.

Il Corpus-based Disambiguator di TMF è il più vicino alla logica di DBpedia Spotlight, con l'aggiunta dei controlli morfo-sintattici che mirano a garantire una maggiore precisione nella distinzione tra le Named Entities (titoli di opere, nomi di persona e di luogo) e i sostantivi comuni.

[161] Lista di stringhe non configurabile dall'utente, in quanto facente parte della logica core di disambiguazione di TMF.

Test No.	Words No.	Total time	Time per word	Precision[162]	Recall[163]
1	449	28 s.	0,06 s.	0,90	0,22
2	110	6 s.	0,05 s.	0,81	0,20
3	115	9 s.	0,07 s.	0.80	0,22
4	139	6 s.	0,04 s.	0,85	0,21
5	146	4 s.	0,02 s.	0,89	0,20
6	127	9 s.	0,07 s.	0,80	0,19
7	174	10 s.	0,05 s.	0,90	0,22
8	150	9 s.	0,04 s.	0,82	0,23
9	205	5 s.	0,03 s.	0,87	0,23
10	147	4 s.	0,02 s.	0,91	0,20
Average	176	9 s.	0,04 s.	0,85	0,21

Tabella 4 - Risultati dell'annotazione corpus-based

I risultati di Tabella 4 indicano una *recall* media alta (il valore 0,21 significa che 21 parole su 100 sono state annotate) e una precisione media discreta (85% di annotazioni corrette), con tempi competitivi (0,04 s. per parola, ovvero 4 secondi per un testo di 100 parole).

Il Corpus-based Disambiguator è la soluzione migliore quando si voglia privilegiare la *recall* pur mantenendo buoni livelli di precisione in tempi rapidi. Può essere indicato per l'annotazione *live* di articoli di giornale o di blog, consentendo magari all'utente di eliminare manualmente le annotazioni ritenute scorrette. Giocando con la *confidence* e col *support* si possono chiaramente ottenere livelli di affidabilità diversi.

[162] In assenza di un gold standard, qui la precisione è stata calcolata come il rapporto tra il numero di annotazione corrette e il numero di annotazioni totali. I valori di confidence e support con cui il test è stato effettuato sono rispettivamente 0.4 e 0.

[163] In assenza di un gold standard, qui la recall è stata calcolata come il rapporto tra il numero di annotazioni e il numero di parole totali.

3.4.2 Knowledge-based disambiguation

La disambiguazione *knowledge-based* utilizza sia il Corpus Index, nella stessa modalità della disambiguazione *corpus-based*, sia il KB Index di TMF, per annotare solo un insieme ristretto di risorse che hanno una marcata affinità concettuale tra loro.

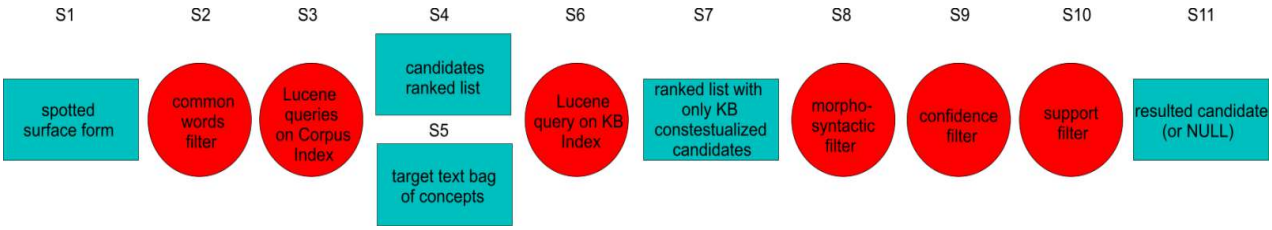


Figura 18 - Processo di annotazione knowledge-based

La surface form individuata in fase di spotting (**S1**) passa inizialmente per il *common words filter* (**S2**), che elimina a priori le parole contenute nel file Common Words (A8). Se passa indenne da questo filtro, la surface form viene utilizzata come elemento costitutivo di due Lucene queries sul Corpus Index (**S3**). La prima è dello stesso tipo della query utilizzata nella disambiguazione Corpus-Based, e produce una *ranked list* dei candidati (**S4**) di quella surface form. La seconda differisce dalla prima in quanto non cerca la surface form nei campi SURFACE FORMS e PREFERRED SURFACE FORM dell'indice, ma esclusivamente va a cercare l'intero testo da annotare nel campo CONTEXTS dell'indice stesso.

Questa seconda query produce un output intermedio molto importante per TMF, ovvero la *bag of concept* del testo target (**S5**). Si tratta di una lista ordinata dei 15 URI che hanno maggiore aderenza contestuale col testo target, e quindi, dei concetti che possono maggiormente essere considerati come l'argomento specifico del testo.

Ognuno dei candidati contenuti in S4 diventa elemento costitutivo di una nuova query Lucene, questa volta lanciata sul KB Index (**S6**), per ottenere dal

campo KB la *bag of concepts* della risorsa candidata[164] e metterla a confronto con la *bag of concepts* del testo target (S5). Solo i candidati che hanno almeno due concetti in comune con S5 non vengono scartati, ottenendo una nuova *ranked list* coi soli candidati "KB-contestualizzati" (**S7**).

S7 viene poi sfoltita, come nella disambiguazione *corpus-based*, dai filtri **S8**, **S9** ed **S10**, fino ad arrivare a un candidato vincente o alla mancata annotazione della *surface form* (**S11**).

Il Knowledge-based Disambiguator è stato progettato con l'intento di annotare i soli concetti fortemente contestualizzati nel documento. Ciò spiega i valori di *recall* molto bassi (0,05, ovvero vengono annotate in media 5 parole su 100), controbilanciati da livelli di precisione vicini ad 1 (circa il 100% delle annotazioni sono corrette).

Test No.	**Words No.**	**Total time**	**Time per word**	**Precision**[165]	**Recall**[166]
1	449	54 s.	0,12 s.	1	0,05
2	110	9 s.	0,08 s.	1	0,05
3	115	12 s.	0,10 s.	1	0,05
4	139	12 s.	0,08 s.	0,90	0,05
5	146	9 s.	0,06 s.	1	0,04
6	127	13 s.	0,10 s.	1	0,05
7	174	14 s.	0,08 s.	1	0,06
8	150	12 s.	0,05 s.	1	0,05
9	205	7 s.	0,04 s.	1	0,05
10	147	8 s.	0,05 s.	1	0,04
Average	176	15 s.	0,07 s.	0,99	0,05

Tabella 5 - Risultati dell'annotazione knowledge-based

[164] La bag of concepts di una risorsa di DBpedia, contenuta nel KB Index, è costituita dall'elenco di tutti gli URI che compaiono almeno due volte come wikilinks nella pagina Wikipedia di riferimento di quella risorsa.

[165] Vedi nota X.

[166] Vedi nota 12

Per questo motivo tale disambiguatore è utile più per la classificazione che per l'annotazione, in particolar modo se consideriamo il valore dell'artefatto intermedio S5 (*target text bag of concept*). Questa lista contiene di fatto, con un grado di affidabilità estremamente alto, i concetti più rilevanti del testo e dunque i principali "argomenti" del documento (anche quando essi non vengano esplicitamente citati nel testo stesso).

Il Knowledge-based Disambiguator è la soluzione migliore quando si voglia privilegiare la precisione a scapito della *recall*, mantenendo tempi ragionevoli (7 secondi per un testo di 100 parole). Può essere utilizzato per classificare automaticamente i documenti di un archivio o di un'attività online in base a tag concettuali connessi a DBpedia, e di conseguenza per l'Ontology-based Information Retrieval.

3.4.3 First Sense Heuristic disambiguation

La disambiguazione di tipo First Sense Heuristic sfrutta il coefficiente di *prominence* delle risorse, ovvero il numero di volte in cui vengono referenziate in Wikipedia attraverso un *wikilink*, per decidere sul significato più comune di un termine ambiguo.

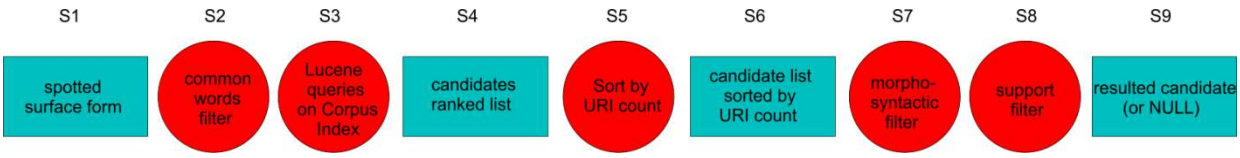


Figura 19 - Processo di annotazione First Sense Heuristic

La *surface form* individuata in fase di *spotting* (**S1**) passa inizialmente per il *common words filter* (**S2**), che elimina a priori le parole contenute nel file Common Words (A8). Se passa indenne da questo filtro, viene utilizzata come elemento costitutivo di una query Lucene sul Corpus Index (**S3**), ottenendo una *ranked list* dei candidati (**S4**), esattamente come accade nella disambiguazione *corpus-based*.

La lista S4 viene poi riordinata (**S5**) secondo il valore del campo URI COUNT di ogni risorsa candidata, dalla risorsa più linkata in Wikipedia a quella meno linkata, producendo la nuova lista di candidati **S6**.

S6 viene sfoltita solo dai filtri **S7** ed **S8** (non c'è più il *confidence filter* in quanto non si vogliono più utilizzare gli *score* ottenuti dalla query Lucene S3), fino ad arrivare a un candidato vincente o alla mancata annotazione della *surface form* (**S9**).

Il First Sense Heuristic Disambiguator è di fatto una "semplificazione" del Corpus-based Disambiguator, in quanto riclassifica i candidati secondo la loro popolarità in Wikipedia anziché secondo la similarità dei contesti.

I tempi di annotazione sono pressoché identici a quelli del Corpus-based Disambiguator, perché il riordino per URI count dei candidati di ogni surface form impiega pochi millesimi di secondo. La *recall* è leggermente più alta (0,24), in quanto, una volta riordinati per URI count, i candidati non sono più filtrati dal *confidence filter*, che si basa sullo *score* dei risultati della query Lucene. I livelli di precisione invece sono più bassi (0,78), come prevedibile per un meccanismo di disambiguazione meno sofisticato.

Test No.	**Words No.**	**Total time**	**Time per word**	**Precision**[167]	**Recall**[168]
1	449	28 s.	0,06 s.	0,85	0,22
2	110	6 s.	0,05 s.	0,80	0,22
3	115	9 s.	0,07 s.	0.70	0,26
4	139	6 s.	0,04 s.	0,76	0,23
5	146	4 s.	0,02 s.	0,79	0,24
6	127	9 s.	0,07 s.	0,80	0,24
7	174	10 s.	0,05 s.	0,85	0,27
8	150	9 s.	0,04 s.	0,71	0,22

[167] Vedi nota X.

[168] Vedi nota 12.

9	205	5 s.	0,03 s.	0,76	0,25
10	147	4 s.	0,02 s.	0,80	0,24
Average	176	9 s.	0,04 s.	0,78	0,24

Tabella 6 - Risultati dell'annotazoione First Sense Heuristic

Lo scenario di utilizzo del First Sense Heuristic Disambiguator può essere analogo a quello del Corpus-based Disambiguator, in ambiti meno specifici e più generici, dove ha più senso suggerire come primo significato di una parola quello più comune in Wikipedia

3.4.4 Default disambiguation

Il Default Disambiguator è stato progettato per ottimizzare le prestazioni del sistema in termini di precisione e *recall*, utilizzando in maniera calibrata gli elementi presenti negli altri disambiguatori e aggiungendo l'ulteriore controllo di alcuni candidati sul Residual KB Index.

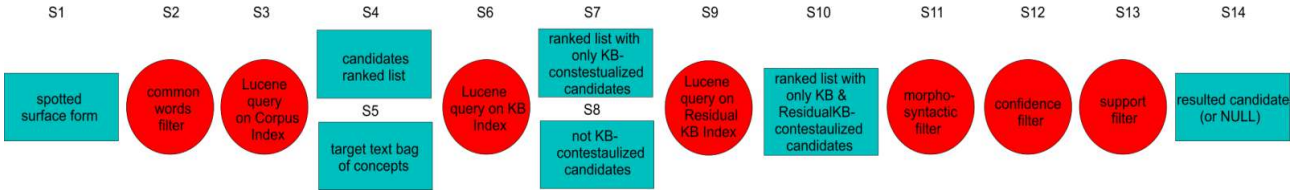


Figura 20 - Proceso di annotazione di default

La *surface form* individuata in fase di *spotting* (**S1**) passa inizialmente per il *common words filter* (**S2**), che elimina a priori le parole contenute nel file Common Words (A8). Se passa indenne da questo filtro, viene utilizzata come elemento costitutivo di due query Lucene sul Corpus Index (**S3**), come nella disambiguazione *knowledge-based*.

Ancora come nella disambiguazione *knowledge-based*, ognuno dei candidati contenuti in S4 diventa elemento costitutivo di una nuova query Lucene lanciata sul KB Index (**S6**), per ottenere dal campo KB la *bag of concepts* della risorsa candidata e metterla a confronto con la *bag of concepts* del testo target (S5). Ne risulta una nuova *ranked list* coi soli candidati "KB-contestualizzati" (**S7**), ma,

contrariamente a quanto avviene nella *KB disambiguation*, anche i candidati non "KB-contestualizzati", ovvero i candidati che sono stati eliminati dal passaggio S6, vengono conservati (**S8**).

Ognuno di essi diventa a sua volta elemento costitutivo di una nuova query Lucene, questa volta lanciata verso il Residual KB Index (**S9**), per ottenere dal campo RESIDUAL KB la *residual bag of concepts* della risorsa candidata e metterla a confronto con la lista delle risorse S7. Solo i candidati la cui *residual bag of concepts* ha almeno un URI in comune con S7 non vengono scartati, dunque ciò che si ottiene alla fine è una nuova *ranked list* coi soli candidati "KB-contestualizzati" e "residualKB-contestualizzati" (**S10**).

Il Residual KB Index contiene gli URI che, nella loro pagina di riferimento in Wikipedia, hanno solo *wikilinks* che appaiono una sola volta nel testo, cioè gli URI le cui pagine Wikipedia non citano mai una stessa risorsa due o più volte. Questo significa che, per quegli specifici URI, non è possibile individuare una effettiva *bag of concepts*, in quanto i *wikilink* che appaiono una sola volta nella pagina non sono affidabili per scoprire il contesto concettuale della risorsa. I *wikilink* associati ad un URI nel Residual KB Index diventano invece utili se vengono messi a confronto coi candidati già filtrati attraverso il procedimento *knowledge-based* (S7), perché le risorse in S7 hanno una sicurezza vicina al 100% e sono risorse effettivamente citate nel testo target (invece la *bag of concepts* del testo target S5 può contenere anche risorse mai citate nel testo). Per questo, anche una sola risorsa in comune con la *residual bag of concepts* di un candidato presente in S8 rende il candidato un probabile risultato positivo[169].

[169] Prendiamo una frase di esempio che abbia come S5 la seguente lista "Letteratura, Giovanni Pascoli, Poesia Italiana" e come S7 la lista "Giovanni Pascoli, Giacomo Leopardi, Dante Alighieri". Le risorse presenti in S7 sono effettivamente presenti nel testo, mentre quelle in S5 possono o meno essere presenti, in quanto costituiscono l'argomento generale del testo. Abbiamo un candidato "Detto d'Amore" (un'opera minore di Dante) che è stato scartato dalla disambiguazione *knowledge-based*, in quanto, avendo una pagina su Wikipedia molto breve, non ha un *wikilink* citato almeno due volte. Ha tuttavia una *residual bag of concepts* dove Dante Alighieri compare una sola volta. Andando a interrogare il Residual KB Index, il Default Disambiguator ottiene la *residual bag of concepts* di "Detto d'Amore", la paragona con S7 e trova la corrispondenza "Dante Alighieri". In questo modo "Detto d'Amore", scartato dalla chiamata al KB Index, viene rimesso in gioco dalla chiamata al Residual KB Index.

S10 viene poi sfoltita, come nella disambiguazione *corpus-based*, dai filtri **S11**, **S12** ed **S13**, fino ad arrivare a un candidato vincente o alla mancata annotazione della *surface form* (**S14**).

La Tabella 7 mostra che il Default Disambiguator, tra gli annotatori di TMF, ha ottenuto i migliori risultati come compromesso tra precisione e *recall*. Infatti affianca ad una precisione del 96% una recall dell'8% (8 parole su 100 vengono annotate) cogliendo, grazie all'utilizzo del KB Index, i concetti semanticamente più interconnessi.

Test No.	**Words No.**	**Total time**	**Time per word**	**Precision**[170]	**Recall**[171]
1	449	79 s.	0,17 s.	0,94	0,08
2	110	11 s.	0,10 s.	0,90	0,09
3	115	17 s.	0,14 s.	1	0,08
4	139	19 s.	0,13 s.	1	0,08
5	146	13 s.	0,08 s.	1	0,08
6	127	19 s.	0,14 s.	1	0,05
7	174	20 s.	0,11 s.	1	0,08
8	150	16 s.	0,07 s.	1	0,07
9	205	10 s.	0,06 s.	0,85	0,08
10	147	12 s.	0,08 s.	0,92	0,09
Average	176	21 s.	0,10 s.	0,96	0,08

Tabella 7 - Risultati dell'annotazione di default

Il Default Disambiguator ha un meccanismo di disambiguazione simile al Knowledge-Based, ma, al fine di aumentare la *recall*, aggiunge ai risultati del Knowledge-Based Disambiguator altre annotazioni "rimesse in gioco" attraverso chiamate al ResidualKB Index. Avendo il KB Disambiguator una precisione vicina al 100%, le annotazioni provenienti dalle query al KB Index vengono utilizza-

[170] Vedi nota 11.

[171] Vedi nota 12.

te come punto di partenza per l'ulteriore disambiguazione di quelle risorse scartate dal KB Disambiguator stesso.

A causa di questo ulteriore *step* nel processo di disambiguazione, i tempi del Default Disambiguator sono leggermente più alti rispetto agli altri (media di 10 secondi per un testo di 100 parole), dunque questo disambiguatore è sconsigliato per l'annotazione *live* di testi. Risulta invece molto utile per l'annotazione offline, per l'estrazione delle principali informazioni dai testi, per la classificazione automatica e per il successivo *enhancement*, i principali obiettivi del progetto TellMeFirst.

3.5 Modulo di classificazione

Il modulo di classificazione del sistema è indipendente da quello di annotazione (*spotting* + disambiguazione) ed espone perciò un servizio REST apposito, chiamato Classify. Tale servizio è finalizzato a reperire, a partire da un testo, i concetti in forma di URI di DBpedia che possono essere considerati gli "argomenti principali" del testo stesso.

3.5.1 Input e output

Classify prende in input quattro parametri[172]:

1) il testo da classificare;
2) il file contenente il testo da classificare;
3) il numero di argomenti che costituiscono la classificazione (10 di default);
4) la lingua.

Il parametro 2 viene letto solo se il parametro 1 è null, altrimenti viene trascurato. Se il parametro 4 è null, il sistema tenta autonomamente la *language detection* con JTCL[173].

L'output del servizio di classificazione è un XML o un JSON[174] contenente una lista di risorse. Ogni elemento di questa lista contiene:

[172] Se non specificato altrimenti, qui come altrove, il default dei parametri è null.
[173] URL: http://textcat.sourceforge.net/

1) l'URI della risorsa di DBpedia corrispondente all'argomento;
2) la label della risorsa di DBpedia nella lingua scelta o identificata automaticamente;
3) lo *score* della risorsa, ovvero il coefficiente che esprime l'affinità di quell'argomento con il testo target.

Le URI al punto 1 sono quelle di DBpedia inglese in tutti i casi, tranne quando non esiste una corrispondente inglese di una risorsa di DBpedia italiana, nel qual caso viene utilizzata la risorsa di DBpedia italiana.

3.5.2 Funzionamento

Il funzionamento del servizio di classificazione è schematizzato in Figura 21.

Il sistema controlla inizialmente se il parametro 1 è null. In tal caso, esegue il *parsing* del file contenuto nel parametro 2 in maniera diversa a seconda della sua estensione (.pdf, .doc, .html, .txt, ecc.). Per l'estrazione del *plain text* dalla struttura del documento vengono utilizzate librerie Java come PDFbox (PDF), Jakarta POI (Word), Snacktory (HTML). Se invece il parametro 1 è diverso da null, il testo viene utilizzato così com'è.

Il testo ricavato dai parametri 1 o 2 viene processato per ottenerne la lunghezza (il numero di parole). Se il numero di parole è inferiore a 1000, l'intero testo viene sottoposto ad analisi da un Lucene Analyzer[175]. Il ruolo dell'Analyzer è quello di tokenizzare il testo, eseguire lo *stemming* ed eliminare le *stop words*, al fine di utilizzare il testo stesso all'interno di una query Lucene booleana sul Corpus Index.

---

[174] Il formato di output dei servizi di TellMeFirst è determinato dal valore di dell'Accept Header della richiesta HTTP. Se la richiesta ha un "Accept: application/xml" il servizio ritorna un XML, se invece ha un "Accept: application/json" il servizio ritorna un JSON. Di default, ovvero senza specificare un Accept Header, il servizio Annotate ritorna un HTML, mentre Classify e tutti i GetX ritornano un JSON.

[175] URL:
http://lucene.apache.org/core/3_6_0/api/all/org/apache/lucene/analysis/Analyzer.html

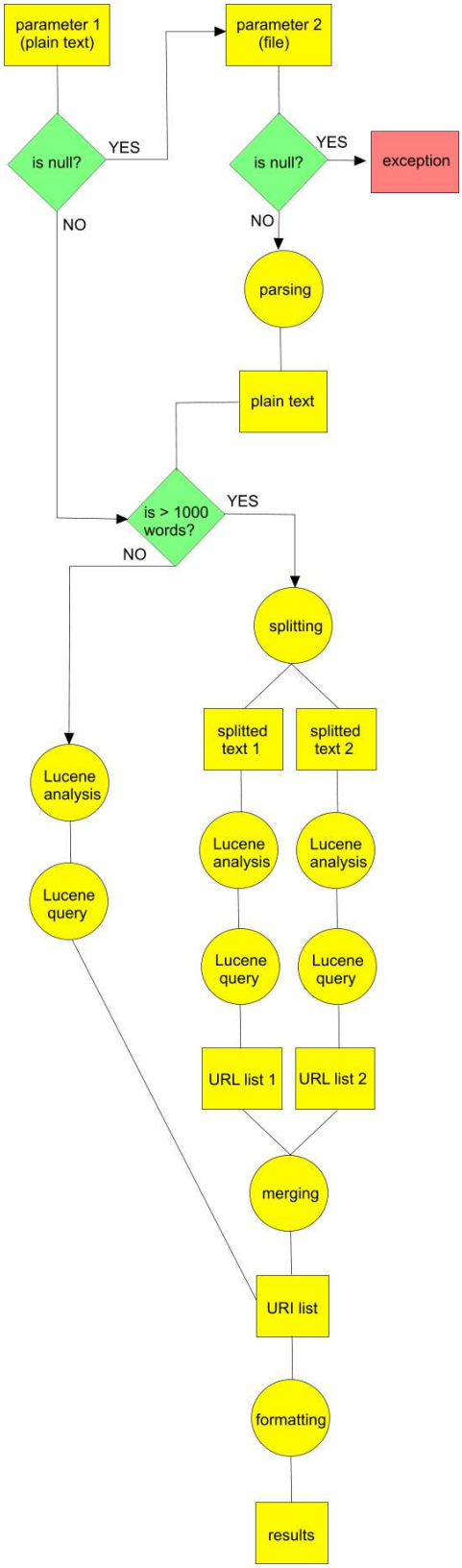


Figura 21 - Schema dell'algoritmo di Classify in TMF

Dato che la query Lucene deve confrontare il testo target col valore del campo CONTEXT di ogni documento Lucene contenuto nel Corpus Index, l'*analyzer* deve essere il medesimo che è stato utilizzato in fase di produzione del Corpus Index. Perciò per l'italiano si usa uno Snowball Analyzer[176] con le *stop words* italiane, mentre per l'inglese uno Standard Analyzer[177]. Per esempio, un testo italiano come "Il giovane regista Roberto Rossellini" viene trasformato nella query booleana sul corpus Index italiano:

```
CONTEXT:giovan OR CONTEXT:regist OR CONTEXT:robert OR CONTEXT:rossellin
```

Da notare che l'articolo "il" è stato eliminato dall'Analyzer in quanto contenuto nella lista delle stopwords italiane, mentre le altre parole sono state sottoposte a *stemming* in base alle regole dello *stemmer* italiano.

La query booleana Lucene ritorna i primi 100 documenti in ordine di *score*. Lo score è calcolato da Lucene stesso in base al coefficiente di similarità di default[178]. Ciò che viene ritornato da Lucene è una lista di documenti, ma, corrispondendo nel Corpus Index ogni documento a una risorsa di DBpedia, è immediato ottenere l'URI della risorsa a partire dal campo URI del documento. A seconda del numero di argomenti richiesto (in input al parametro 3 del servizio

[176] URL:
http://lucene.apache.org/core/3_6_0/api/all/org/apache/lucene/analysis/snowball/SnowballAnalyzer.html

[177] URL:
http://lucene.apache.org/core/3_6_0/api/core/org/apache/lucene/analysis/standard/StandardAnalyzer.html. La differenza tra lo Snowball Analyzer e lo Standard Analyzer è nel meccanismo di *stemming*. Lo Standard Analyzer utilizza l'algoritmo di *stemming* di Martin Porter ottimizzato per la lingua inglese, dove sono tenuti in considerazione, oltre alle regole radice-desinenza dell'inglese, gli acronimi, gli indirizzi mail, i nomi di azienda, i numeri, le parole con un apostrofo interno, ecc. Per le altre lingue europee, esiste una classe ombrello, lo Snowball Analyzer, che prende nel costruttore lo *stemmer* di una particolare lingua (anche customizzato) e che segue le regole di *stemming* per la specifica lingua. Di default Lucene fornisce alcuni *stemmer* che si possono passare nel costruttore dello Snowball indicandone soltanto la lingua: italiano, danese, olandese, finlandese, francese, tedesco, norvegese, portoghese, russo, spagnolo e svedese.

[178] URL:
https://lucene.apache.org/core/3_6_0/api/all/org/apache/lucene/search/DefaultSimilarity.html

Classify) vengono prelevati i primi n URI e inseriti nella lista in output al servizio, insieme alle rispettive label[179] e ai rispettivi score.

Se invece il numero di parole del testo è superiore a 1000, il testo viene suddiviso in sottoinsiemi di massimo 1000 parole[180]. Ogni sottoinsieme viene sottoposto ad analisi dal Lucene Analyzer e utilizzato in una query sul Corpus Index come visto in precedenza. Le analisi e le query per i singoli pezzi avvengono in parallelo, all'interno di n *thread* Java che partono simultaneamente. Quando tutti i *thread* si sono conclusi, i risultati ottenuti dai vari sottoinsiemi vengono mergiati in un'unica lista ed ordinati nel modo descritto di seguito. I risultati che occorrono più di una volta vengono ordinati 1) per numero di occorrenze, 2) per score Lucene. I risultati che occorrono una sola volta vengono ordinati secondo il loro score e posizionati dopo i risultati che occorrono più di una volta. I primi n URI di questa lista unica, in base al numero settato al parametro 3 in input, vengono ritornati dal servizio.

3.5.3 Esempio di chiamata e risposta

Ecco un esempio di chiamata al servizio Classify di TellMeFirst:

```
curl -i -X POST \
-H "Accept:application/json" \
-H «content-type:application/x-www-form-urlencoded" \
-d
«text=Roberto+Rossellini+è+un+regista+molto+conosciuto+all'estero.&numTo
pics=5&lang=italian» \
http://example.com:2222/rest/classify/
```

[179] Il mapping tra label e URI è contenuto nei file
http://downloads.dbpedia.org/3.7/en/labels_en.nt.bz2 (per l'inglese) e
http://downloads.dbpedia.org/3.7-i18n/it/labels_it.nt.bz2 (per l'italiano). Per velocizzare le operazioni, TMF conserva il contenuto di questi file nello stesso database MySQL usato per l'URI mapping.

[180] Attualmente il troncamento del testo avviene mantenendo l'integrità a livello di parola, ma non di frase. Sulla base di alcuni test, il troncamento del testo alla fine della frase ha dato i medesimi risultati del troncamento basato sul numero di parole, perciò si è scelto di non implementarlo per non appesantire ulteriormente l'algoritmo. Questa invarianza accade probabilmente perché la query Lucene elimina congiunzioni, proposizioni e articoli (*stop words*), cercando i *token* nell'indice in modo svincolato dalle loro peculiarità morfosintattiche, e dunque rende trascurabile l'aggiunta o la sottrazione di poche parole rispetto al blocco dei 1000 *tokens*.

La risposta alla chiamata sopra è questa:


```
{
  "Resources":   [
        {
      "uri": "http://dbpedia.org/resource/Roberto_Rossellini",
      "label": "Roberto Rossellini",
      "title": "Roberto Rossellini",
      "score": "0.6799243",
      "mergedTypes":"foaf:Person#dbpedia-owl:Person#dbpedia-
owl:Agent#http://schema.org/Person#yago:PeopleFromRome(city)#yago:Italia
nFilmDirectors#yago:ItalianAtheists#yago:Person100007846"
    },
        {
      "uri": "http://dbpedia.org/resource/Pais%C3%A0",
      "label": "Paisà",
      "title": "Paisà",
      "score": "0.60960734",
      "mergedTypes":"dbpedia-owl:Work#dbpedia-
owl:Film#http://schema.org/Movie#http://schema.org/CreativeWork#yago:Ita
lianFilms#yago:AnthologyFilms#yago:ItalianNeorealistFilms#yago:Black-
and-whiteFilms#yago:1946Films#yago:Italian-
languAgeFilms#yago:FilmsDirectedByRobertoRossellini"
    },
        {
      "uri": " http:dbpedia.org/resource/A_Pilot_Returns",
      "label": "Un pilota ritorna",
      "title": "A Pilot Returns",
      "score": "0.5915979",
      "mergedTypes":"dbpedia-owl:Work#dbpedia-
owl:Film#http://schema.org/Movie#http://schema.org/CreativeWork"
    },
        {
      "uri":
"http://dbpedia.org/resource/Renzo_Rossellini_%28producer%29",
      "label": "Renzo Rossellini",
      "title": "Renzo Rossellini (producer)",
      "score": "0.5777477",
      "mergedTypes":"foaf:Person#dbpedia-owl:Person#dbpedia-
owl:Agent#http://schema.org/Person"
    },
        {
      "uri": "http://dbpedia.org/resource/Isabella_Rossellini",
      "label": "Isabella Rossellini",
      "title": "Isabella Rossellini",
      "score": "0.5492195",
```

```
      "mergedTypes": "foaf:Person#dbpedia-owl:Person#dbpedia-
owl:Agent#http://schema.org/Person#yago:Actor109765278#yago:NaturalizedC
itizensOfTheUnitedStates#yago:LivingPeople#yago:PeopleFromRome(city)#yag
o:Writer110794014#yago:AmericanPeopleOfItalianDescent#yago:TwinPeople#ya
go:ItalianImmigrantsToTheUnitedStates#yago:AmericanActorsOfSwedishDescen
t#yago:ItalianPeopleOfGermanDescent#yago:Person100007846#yago:SwedishPeo
ple#yago:FilmMaker110088390#yago:Philanthropist110421956#yago:ItalianPeo
pleOfSwedishDescent#yago:ItalianFemaleModels#http://umbel.org/umbel/rc/A
ctor#yago:Model110324560yago:ItalianFilmActors"
    }
  ]
}
```


3.6. Modulo di enhancement

Il modulo di *enhancement* di TMF si occupa di reperire informazioni e contenuti aggiuntivi sulle risorse fornite in output dal modulo di classificazione. Espone cinque diversi servizi: GetImage, GetText, GetNews, GetMap e GetVideo. I primi tre servizi vengono chiamati dalla GUI di TellMeFirst indipendentemente dal tipo della risorsa in uscita da Classify, mentre GetMap e GetVideo sono chiamati solo per le risorse che hanno determinati DBpedia *types*[181]. Il risultato di questa scelta è che la GUI caricherà in ogni caso, in prima battuta, un'immagine, un testo e un'insieme di news (se disponibili) per ogni "argomento" del testo, mentre solo in seguito, per determinati DBpedia *types*, caricherà gli altri contenuti, ovvero le mappe e i video.

3.6.1 GetImage

GetImage prende in input i seguenti parametri:

1) l'URI della risorsa di cui reperire un'immagine.
2) la label della risorsa

L'output di GetImage è il percorso completo di un'immagine reperibile sul Web.Il funzionamento del servizio getImage è schematizzato in Figura 22.

[181] Si vedano gli specifici paragrafi dedicati ai servizi.

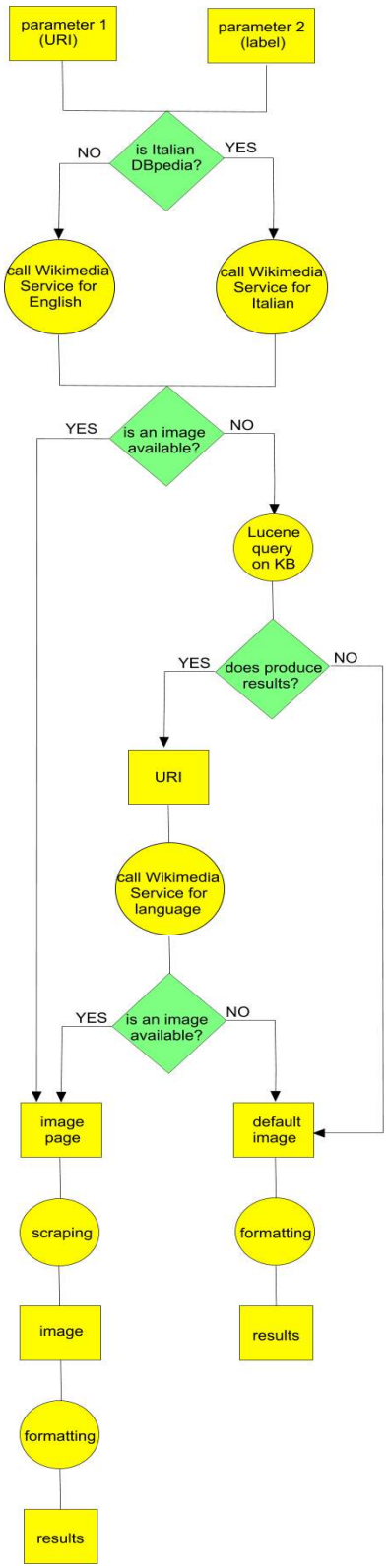


Figura 22 – Schema dell'algoritmo di GetImage in TMF

In primo luogo viene analizzato l'URI in input al parametro 1 per comprendere se si tratta di una risorsa di DBpedia inglese o di DBpedia italiana. Di seguito si prende la label in input al parametro 2 per produrre una richiesta al servizio Images delle API di Wikimedia Commons. Per le risorse di DBpedia inglese si produce una richiesta del tipo: http://en.wikipedia.org/w/api.php?action=query&titles=<label>&prop=images, mentre per le risorse di DBpedia italiana si produce una richiesta del tipo: http://it.wikipedia.org/w/api.php?action=query&titles=<label>&prop=images.

La label contenuta nel parametro 2 corrisponde al titolo dell'articolo di Wikipedia per la risorsa specificata nel parametro 1, perché è ottenuta dall'output del servizio Classify: il servizio Images delle API di Wikimedia si aspetta proprio il titolo di una pagina Wikipedia[182].

La risposta del servizio Images è un XML di questo tipo:

```
<api>
  <query>
    <pages>
      <page pageid="736" ns="0" title="Albert Einstein">
        <images>
          <im ns="6" title="Image:1919 eclipse positive.jpg" />
          <im ns="6" title="Image:Albert Einstein Head.jpg" />
          <im ns="6" title="Image:Albert Einstein german.ogg" />
          <im ns="6" title="Image:Albert Einstein photo 1921.jpg" />
          ...
        </images>
      </page>
    </pages>
  </query>
</api>
```

L'XML viene parsificato per ottenere la lista contenente i nomi delle immagini. Viene dunque applicato l'algoritmo Jaro-Winkler[183], implementato dalla classe JaroWinklerDistance di LingPipe[184], per rinvenire il nome più simile alla label

[182] È utile ricordare che titoli delle pagine di Wikipedia per ogni lingua sono univoci in base alla politica editoriale di Wikipedia.

[183] URL: http://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

[184] URL: http://alias-i.com/lingpipe/docs/api/com/aliasi/spell/JaroWinklerDistance.html

della risorsa. Una volta ottenuto, vi si aggiunge il prefisso http://it.wikipedia.org/wiki/File: per ottenere la pagina di Wikimedia nella quale è embeddata l'immagine. Quindi si esegue lo *scraping* della pagina di Wikimedia per ottenere il percorso reale dell'immagine, che viene passato in output. Se con questo procedimento non si ottiene alcuna immagine, per esempio perché Wikimedia non dispone di alcuna immagine per quella particolare risorsa, viene lanciata una query Lucene sul KB Index per ricevere la *bag of concetps* della risorsa, al fine di ottenere l'URI di un argomento affine a quello desiderato. Se la risorsa non dispone di una *bag of concepts*, oppure se nessuno degli URI della *bag of concepts* produce un risultato dal servizio Images della API di Wikimedia, allora il sistema restituisce in output il percorso di un'immagine di default salvata sul server.

Ecco un esempio di chiamata al servizio getImage di TellMeFirst:

```
curl
"http://example.com/rest/getImage?uri=http://dbpedia.org/resource/Robert
o_Rossellini&label=Roberto+Rossellini" -H "Accept:application/json"
```

La risposta a tale chiamata è la seguente:

```
[
 {
  "imageURL":
"http://upload.wikimedia.org/wikipedia/en/0/09/Roberto_Rossellini.jpg"
 }
]
```

3.6.2 GetText

Il servizio getText ha come unico parametro in ingresso l'URI della risorsa per cui si vuole ottenere una descrizione testuale. In uscita al servizio abbiamo una stringa contenente l'URL della pagina Wikipedia che si riferisce alla specifica risorsa. La GUI effettuerà lo *scraping* di tale pagina per presentare in maniera adeguata il corpo, i titoli, i sottotitoli, le note, la bibliografia, ecc.

A partire dall'URI in ingresso, il sistema ottiene il titolo della pagina Wikipedia della risorsa, che verrà in seguito utilizzata per ottenere il testo della pagi-

na di Wikipedia a cui appartiene quel titolo[185]. L'URL della pagina è ritornato in output. Ecco un esempio di chiamata al servizio getText di TellMeFirst:

```
curl
"http://example.com/rest/getText?uri=http://dbpedia.org/resource/Giorgio
_Armani" -H "Accept:application/json"
```

La risposta a tale chiamata è la seguente:

```
[
 {
  "title": "Giorgio Armani"
 }
]
```

3.6.3 GetNews

Il servizio GetNews ha come unico parametro in ingresso l'URI della risorsa per la quale si vogliono visualizzare delle notizie di giornale. In uscita al servizio abbiamo un XML o un JSON contenente gli *snippet* delle notizie che il New York Times fornisce per mezzo delle sue API[186].

Il sistema controlla se l'URI in input appartiene a una risorsa DBpedia italiana o inglese. Nel primo caso, il valore ritornato in uscita dal servizio è null, in quanto solo le risorse di DBpedia inglese hanno un corrispondente nei New York Times Linked Open Data (NYTLOD)[187]. Appurato che si tratti di una risorsa di DBpedia inglese, il sistema controlla che abbia un corrispondente *sameAs* tra le risorse in NYTLOD[188], facendosi restituire l'URI della risorsa di NYTLOD.

---

[185] La corrispondenza tra URI e titoli delle pagine Wikipedia è contenuta nei file http://downloads.dbpedia.org/3.7/en/labels_en.nt.bz2 (inglese) e http://downloads.dbpedia.org/3.7-i18n/it/labels_it.nt.bz2 (italiano). Naturalmente è possibile ottenere questa corrispondenza anche via query SPARQL sull'endpoint di DBpedia e di DBpedia Italia, tuttavia per velocizzare le operazioni TMF conserva il contenuto di questi file nello stesso database MySQL usato per l'URI mapping.

[186] URL: http://developer.nytimes.com/

[187] URL: http://data.nytimes.com/

[188] La corrispondenza tra URI di DBpedia e URI di NYTLOD è contenuta nel file http://wiki.dbpedia.org/Downloads37#linkstonewyorktimes. Naturalmente è possibile ottenere questa corrispondenza anche via query SPARQL sull'endpoint di DBpedia (cercando tra i valori della proprietà "owl:sameAs"), tuttavia per velocizzare le operazioni TMF conserva il contenuto di questo file nello stesso database MySQL usato per l'URI mapping.

TellMeFirst chiede al server del New York Times la dereferenziazione di questo URI specificando nell'HTTP Accept Header il valore "application/rdf\+xml". In tal modo ottiene una descrizione in RDF della risorsa, dalla quale può estrarre il valore della proprietà *nyt:search_api_query*. Questo valore è di fatto una richiesta al servizio REST Search del NYT per reperire gli articoli più recenti apparsi sulla testata online per l'argomento specificato. Viene effettuata la chiamata al servizio, che ritorna un JSON.

Ecco un esempio di chiamata al servizio getNews di TellMeFirst:

```
curl
"http://example.com/rest/getNews?uri=http://dbpedia.org/resource/Afghani
stan" -H "Accept:application/json"
```

Un esempio di risposta in JSON è la seguente, in cui sono contenuti i ritagli di due notizie riguardanti l'argomento "Afghanistan":

```
{
     "results":[
      {
         "body":"The most important line in President Obama's Afghan
speech was not about Afpak policy (so named by the White House) but
about the U.S. domestic situation: "Our troop commitment in Afghanistan
cannot be open-ended - because the nation that I am most interested in
building is our own." As military strategy for winning a war",
         "date":"20091208",
         "des_facet":[
            "UNITED STATES DEFENSE AND MILITARY FORCES",
            "UNITED STATES ECONOMY",
            "AFGHANISTAN WAR (2001- )"
         ],
         "title":"OP-ED COLUMNIST; Afghanistan on Main Street",
         "url":"http:\/\/www.nytimes.com\/2009\/12\/08\/opinion\/08iht-
edcohen.html"
      },
      {
         "body":"The top military commander in Afghanistan warns in a
confidential assessment of the war there that he needs additional troops
within the next year or else the conflict ''will likely result in
failure.'' The grim assessment is contained in a 66-page report that the
commander, Gen. Stanley A. McChrystal, submitted to Defense Secretary
Robert M. Gates",
         "date":"20090921",
         "des_facet":[
            "UNITED STATES DEFENSE AND MILITARY FORCES",
```

```
            "AFGHANISTAN WAR (2001- )"
         ],
         "title":"General Calls for More Troops To Avoid Afghanistan
Failure",

"url":"http:\/\/www.nytimes.com\/2009\/09\/21\/world\/asia\/21afghan.htm
l"
      }
   ]
}
```


Come si può osservare, i campi dell'articolo sono facilmente identificabili: *body*, *date*, *des_facet*[189], *title*, *url*. La GUI utilizzerà questi campi per formattare il ritaglio dell'articolo e rimandare l'utente alla pagina completa dell'articolo attraverso un link "Read full text".

3.6.4 GetMap

Il servizio GetMap ha come unico parametro in ingresso l'URI della risorsa che si vuole rappresentare su mappa. In uscita al servizio abbiamo un XML (o un JSON) contenente i valori di latitudine e longitudine del luogo richiesto.

La GUI di TellMeFirst chiama il servizio GetMap solo per le risorse che hanno tra i tipi *dbpedia-owl:Place*. Il sistema chiama lo SPARQL endpoint di DBpedia per ottenere le proprietà *geo:lat* (latitudine) e *geo:long* (longitudine) della risorsa di tipo Place, dunque le passa in uscita al servizio.

Ecco un esempio di chiamata al servizio getMap di TellMeFirst:

```
curl
"http://example.com/rest/getMap?uri=http://dbpedia.org/resource/Italy" -
H "Accept:application/json"
```

La risposta a tale chiamata è la seguente:


```
[{
  "lat": "41.900002",
  "long" : "12.483334"
 }]
```


[189] Sono i tag, o *facets*, dell'articolo.

3.6.5 GetVideo

Il servizio GetVideo prende in input i seguenti parametri:

1) l'URI della risorsa di cui reperire un video;
2) la label della risorsa;
3) il DBpedia type della risorsa.

In uscita al servizio abbiamo l'URL di un video di YouTube.

La GUI di TellMeFirst chiama il servizio GetVideo solo per le risorse che hanno tra i tipi *dbpedia-owl:Actor*, *dbpedia-owl:Activity*, *dbpedia-owl:Band*, *dbpedia-owl:Artist* (escluso *dbpedia-owl:Writer*), *dbpedia-owl:Athlete*, o *dbpedia-owl:MusicalWork*. Questa scelta rappresenta un compromesso tra il possibile interesse dell'utente nei confronti del video di una risorsa e l'incertezza nell'affidabilità dei risultati (YouTube non ha meccanismi di tipo semantico per disambiguare i tag delle ricerche).

Il sistema verifica il parametro 3 e si comporta in maniera differente a seconda dei seguenti casi:

1) il parametro 3 è null. Il sistema ricava il *type* della risorsa a partire dal parametro 1 per mezzo di una query Lucene sul Corpus Index (italiano o inglese a seconda del dominio dell'URI). Se nella lista dei *type* non è contenuto uno tra *dbpedia-owl:Actor*, *dbpedia-owl:Activity*, o *dbpedia-owl:Band*, o *dbpedia-owl:Artist* (escluso *dbpedia-owl:Writer*), o *dbpedia-owl:Athlete*, o *dbpedia-owl:MusicalWork*, allora il servizio ritorna null. Al contrario, se il *type* è *dbpedia-owl:MusicalWork*, si veda il punto 2; se il *type* non è *dbpedia-owl:MusicalWork* si veda il punto 3.
2) il *type* è *dbpedia-owl:MusicalWork*. TellMeFirst controlla se l'URI in input al parametro 1 appartiene a una risorsa di DBpedia italiana o inglese. Quindi lancia una chiamata sull'endpoint SPARQL di DBpedia (italiana o inglese) per ottenere la label del valore della proprietà *dbpprop:artist* dell'URI contenuto nel parametro 1. Una volta ottenuto questa label, produce per concatenamento una stringa del tipo: <label dell'artista> + spazio + <label contenuta nel parametro 2>[190]. Questa scelta è motivata dal fatto che, aggiungendo l'autore dell'opera musicale, si hanno minori

[190] Esempio: "Ludovico Einaudi Una Mattina".

probabilità che la richiesta sia ambigua. La suddetta stringa viene utilizzata all'interno di una richiesta al servizio REST Search di YouTube, come nell'esempio: https://gdata.youtube.com/feeds/api/videos?q=Ludovico+Einaudi+Una+Mattina&start-index=1&max-results=1&v=2&hd=true
Il parametro "hd" è impostato a "true" per cercare di ottenere i video di maggiore qualità

3) il type non è *dbpedia-owl:MusicalWork*. La label del parametro di input 2 viene utilizzata all'interno di una richiesta al servizio REST Search di YouTube, come nell'esempio: https://gdata.youtube.com/feeds/api/videos?q=Ludovico+Einaudi&start-index=1&max-results=1&v=2&hd=true
Anche in questo caso il parametro "hd" è impostato a "true" per cercare di ottenere i video di maggiore qualità.

Ecco un esempio di chiamata al servizio getVideo di TellMeFirst:

```
curl
"http://example.com/rest/getVideo?uri=http://dbpedia.org/resource/Ludovi
co_Einuadi &label=Ludovico+Einaudi&type=dbpedia-owl:Artist" -H
"Accept:application/json"
```

La risposta alla chiamata sopra è questa:

```
[
 {
  "videoURL": "http://youtu.be/hHsnECVc_DE"
 }
]
```

3.7 Interfaccia Web

La GUI di TellMeFirst è un'applicazione Web scritta in Javascript (JQuery), HTML e CSS. Durante l'utilizzo propone all'utente tre principali elementi interattivi: il modulo di acquisizione del testo, la griglia degli argomenti e il box dei contenuti. Di seguito sono esposte singolarmente le caratteristiche di questi oggetti.

3.7.1 Modulo di acquisizione del testo

Il primo elemento è un *form* per l'input del testo nel sistema (Figura 23). Le modalità di inserimento del testo sono tre:

1) digitazione del testo in una *text area*;
2) digitazione di un URL in una *text field*;
3) caricamento di un file attraverso un *file chooser* (PDF, DOC o TXT).

Figura 23 - Modulo di acquisizione del testo

Questi tre componenti del *form* sono nascosti dietro le voci di un menu ("Upload a file", "Input a text" e "Insert the URL of an article") e compaiono espandendo le singole voci senza che la pagina venga ricaricata. La voce "Input a text" compare di default già aperta. Al di sotto di questo menu vi è un *radio button* composto da due elementi a forma rispettivamente di bandiera italiana e inglese, per la scelta della lingua del testo. Di default è selezionata la bandiera inglese. Nell'angolo in basso a destra c'è un pulsante per il *submit*. Quando l'utente ha inserito il testo in una delle tre modalità e ha scelto la lingua, premendo il pulsante di *submit* visualizza un'icona di caricamento (*loading wheel*). Quando il

caricamento è stato completato, in luogo del *form* appare il secondo elemento della GUI, ovvero la griglia degli argomenti.

3.7.2 Griglia degli argomenti

La griglia degli argomenti suddivide l'intero schermo in 7 *frames* rettangolari di diversa grandezza (Figure 24 e 25). Il *frame* numero 1, il più grande, è posizionato nell'angolo in alto a sinistra, e ha un'altezza che occupa il 50% dello schermo e una larghezza che ne occupa il 66.6%. I *frame* numero 2 e 3 si dividono equamente la porzione di schermo sotto il numero 1 (altezza 50% dello schermo e larghezza 33.3% ciascuno). Sul lato in alto a destra troviamo i frame 4 e 5, che hanno entrambi altezza 33,3% dello schermo e larghezza 33,5%. In basso a destra ci sono i frame 6 e 7 (altezza 33.6% dello schermo e larghezza 16.8%).

Ogni *frame* indica un argomento estratto dal testo e la sua dimensione ne rappresenta la rilevanza. Ciò significa che gli argomenti sono disposti in ordine decrescente di importanza dalla posizione 1 fino alla 7. In ogni frame è contenuta un'immagine e un'etichetta che si riferiscono all'argomento estratto dal testo.

Figura 24 - Griglia degli argomenti

Al passaggio del mouse su ogni frame, l'immagine si ingrandisce lievemente, con un effetto 3D, e l'etichetta aumenta di luminosità. Cliccando su uno dei frame si visualizza il terzo elemento della GUI, ovvero il box dei contenuti.

3.7.3 Box dei contenuti

Il box dei contenuti è un rettangolo di grandezza variabile che mostra i risultati dell'arricchimento del singolo argomento. Per ogni argomento viene mostrata in sequenza un'immagine tratta da Wikimedia Commons[191], un testo tratto da Wikipedia, i principali metadati tratti da DBpedia, un elenco di notizie tratte dal New York Times (per gli argomenti che ne dispongono), una mappa da OpenStreetMap (per i luoghi) e un video da YouTube (per artisti, personaggi dello spettacolo o celebrità). Il box dispone di due freccette ai suoi lati per consentire all'utente di passare da un contenuto all'altro. La dimensione del box varia a seconda della grandezza del contenuto da visualizzare. Sotto il box è riportata l'etichetta dell'argomento, mentre sopra vi è un'icona che indica la fonte del contenuto. La griglia di partenza rimane in secondo piano rispetto al box, opacizzata dall'effetto di un *layer* scuro. Le Figure da 26 a 31 mostrano la visualizzazione dei diversi contenuti all'interno del box.

La visualizzazione dei metadati utilizzata in TMF (Figura 28) si caratterizza per una struttura di tipo *node-link* a raggiera, in cui il nodo centrale costituisce la risorsa di partenza, mentre gli estremi rappresentano i nodi in cui si collocano le informazioni (i metadati) che sono state recuperate grazie alle query SPARQL: su ciascun nodo viene visualizzata la label recuperata tramite queste stesse query. Qualora i nodi rappresentino specifiche entità, e non soltanto valori di tipo *literal*, un click su ognuno di essi apre una nuova finestra del browser che reindirizza l'utente verso la risorsa corrispondente in DBpedia o in DBpedia Italia. I rami che collegano il nodo centrale ai nodi periferici sono anch'essi opportunamente etichettati (e a loro volta cliccabili in quanto URI) per mostrare all'utente la natura della relazione che collega l'entità centrale con ognuna di quelle

[191] Nel Box l'immagine è visualizzata nelle sue dimensioni originali se è più piccola dello schermo, altrimenti alla massima dimensione consentita dallo schermo. All'interno della griglia, invece, si adatta alla grandezza del proprio frame.

periferiche e di conseguenza completare la visualizzazione delle triple RDF. Poiché la quantità di relazioni che riguardano una specifica entità può essere molto elevata, per non appesantire la visualizzazione è stata adottata una strategia per la quale vengono mostrati soltanto sei nodi alla volta: un primo bottone consente di esplorare i nodi che non sono presenti al momento aggiungendoli alla visualizzazione e rimuovendo i precedenti, mentre un secondo bottone consente di compiere l'operazione a ritroso, per poter mostrare i nodi precedenti.

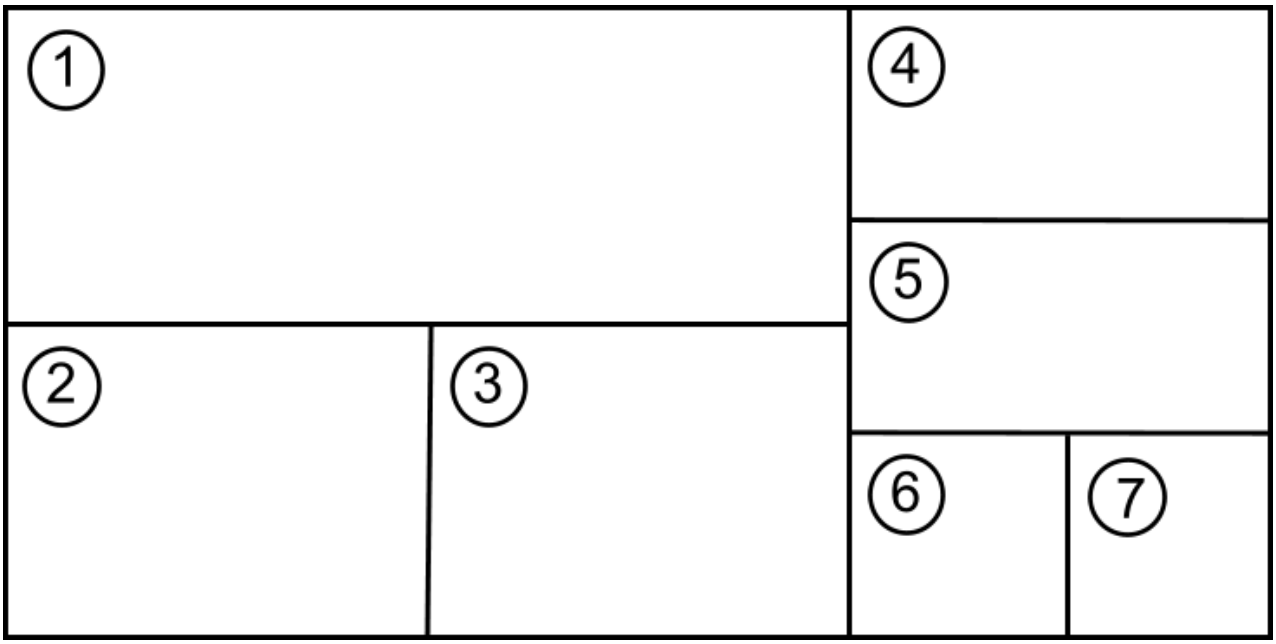


Figura 25 - Griglia degli argomenti (suddivisione dello schermo)

Figura 26 - Box dei contenuti (Immagine)

Figura 27 - Box dei contenuti (testo)

Figura 28 - Box dei contenuti (metadati)

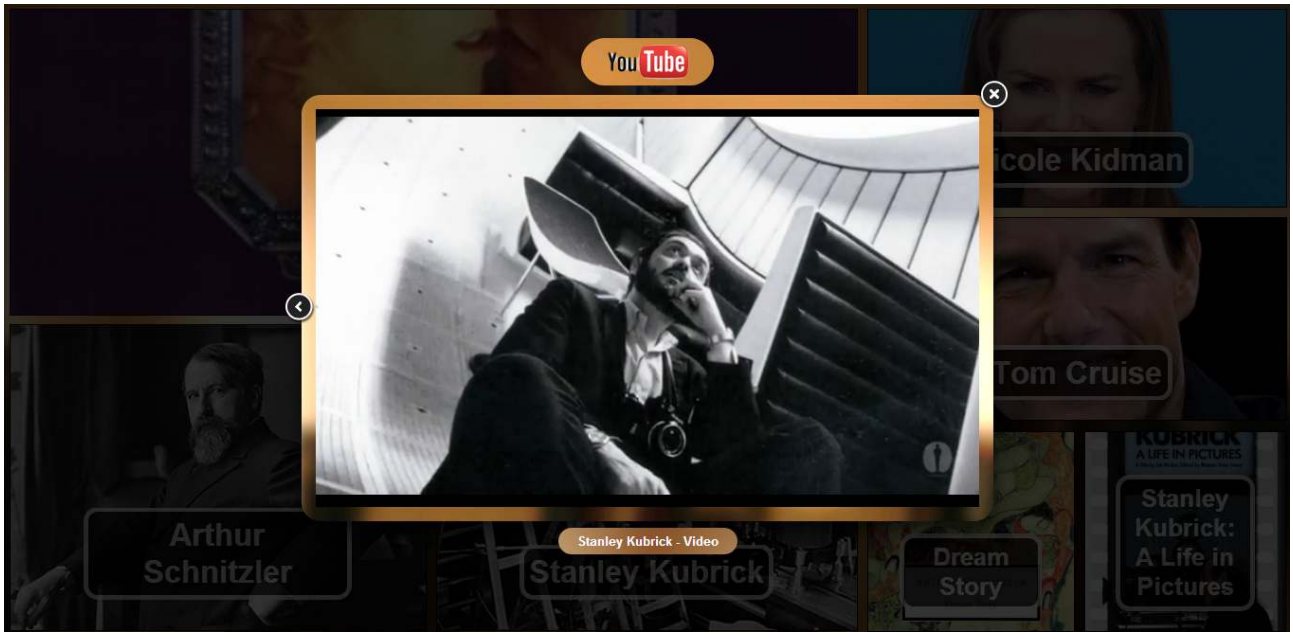


Figura 29 - Box dei contenuti (video)

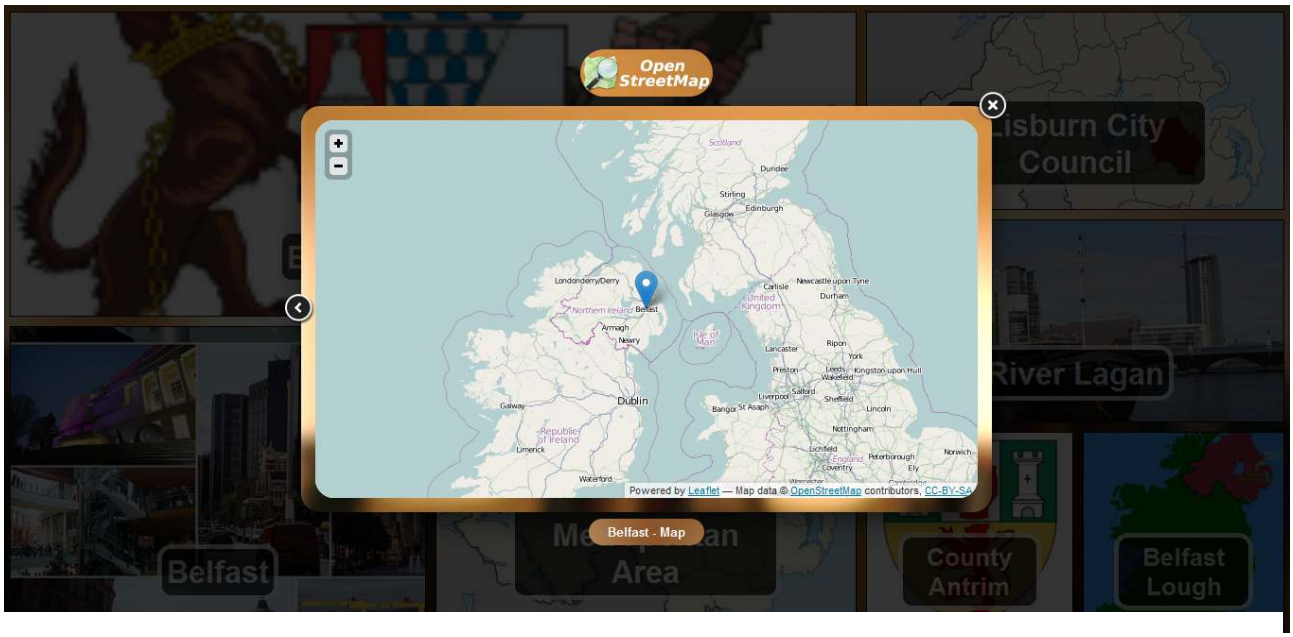


Figura 30 - Box dei contenuti (mappa)

Figura 31 - Box dei contenuti (news)

3.7.4 Flusso di esecuzione

Il flusso di esecuzione di TMF dall'acquisizione del testo alla visualizzazione dei contenuti di *enhancement* è esposto nei paragrafi seguenti ed illustrato sinteticamente in Figura 32.

3.7.4.1 Acquisizione e parsing del testo

Il primo passo, dopo la sottomissione del *form* da parte dell'utente, è il controllo lato client di alcuni valori riguardanti l'input (la lunghezza del testo, l'esistenza della pagina inserita nell'URL, la dimensione e il formato del file caricato, ecc).

Se la richiesta passa questi controlli, l'input viene estratto dal *form* nelle seguenti modalità.

1) Nel caso di inserimento diretto del testo nella *text area*, il testo viene passato così com'è al livello successivo.
2) Nel caso di inserimento di una URL nella *text field*, il sistema utilizza la libreria Snacktory per estrarre il testo principale dalla pagina in questione. Questa libreria, come altre analoghe (Goose[192], Readability[193], ecc), è in grado di isolare la porzione di testo più estesa all'interno della pagina, scartando gli elementi di contorno. La scelta è ricaduta su Snacktory in quanto si è rivelata più precisa nello *scraping* degli articoli di giornale in lingua italiana, con performance comparabili con le altre per diversi tipi di contenuto.
3) Nel caso di caricamento di un file, il sistema utilizza la libreria Apache PDFBox per effettuare l'estrazione del testo dai PDF, la libreria Apache POI per l'estrazione da file Microsoft Word e le librerie standard di java.io per i file di testo semplici (.txt). Sono stati scelti dunque gli

[192] URL: https://github.com/jiminoc/goose/wiki
[193] URL: http://www.readability.com/developers/api

strumenti più noti e utilizzati in questo ambito, che hanno dato i risultati attesi.

3.7.4.2 Chiamate ai servizi

Classify

A partire dal *plain text* ottenuto nel passo precedente, la *webapp* di TellMeFirst chiama il servizio Classify di TMF server, passando come parametri il testo e la lingua scelta. Riceve in risposta un JSON con la seguente struttura:


```
{
  "Resources":   [
        {
      "uri": "http://dbpedia.org/resource/Roberto_Rossellini",
      "label": "Roberto Rossellini",
      "title": "Roberto Rossellini",
      "score": "0.6799243",
      "mergedTypes":"foaf:Person#dbpedia-owl:Person#dbpedia-
owl:Agent#http://schema.org/Person#yago:PeopleFromRome(city)#yago:Italia
nFilmDirectors#yago:ItalianAtheists#yago:Person100007846"
    },
        {
      "uri": "http://dbpedia.org/resource/Pais%C3%A0",
      "label": "Paisà",
      "title": "Paisà",
      "score": "0.60960734",
      "mergedTypes":"dbpedia-owl:Work#dbpedia-
owl:Film#http://schema.org/Movie#http://schema.org/CreativeWork#yago:Ita
lianFilms#yago:AnthologyFilms#yago:ItalianNeorealistFilms#yago:Black-
and-whiteFilms#yago:1946Films#yago:Italian-
languAgeFilms#yago:FilmsDirectedByRobertoRossellini"
    },
        {
      "uri": "http:dbpedia.org/resource/A_Pilot_Returns",
      "label": "Un pilota ritorna",
      "title": "A Pilot Returns",
      "score": "0.5915979",
      "mergedTypes":"dbpedia-owl:Work#dbpedia-
owl:Film#http://schema.org/Movie#http://schema.org/CreativeWork"
    },
        {
      "uri":
"http://dbpedia.org/resource/Renzo_Rossellini_%28producer%29",
```

```
      "label": "Renzo Rossellini",
      "title": "Renzo Rossellini (producer)",
      "score": "0.5777477",
      "mergedTypes":"foaf:Person#dbpedia-owl:Person#dbpedia-
owl:Agent#http://schema.org/Person"
    },
        {
      "uri": "http://dbpedia.org/resource/Isabella_Rossellini",
      "label": "Isabella Rossellini",
      "title": "Isabella Rossellini",
      "score": "0.5492195",
      "mergedTypes": "foaf:Person#dbpedia-owl:Person#dbpedia-
owl:Agent#http://schema.org/Person#yago:Actor109765278#yago:NaturalizedC
itizensOfTheUnitedStates#yago:LivingPeople#yago:PeopleFromRome(city)#yag
o:Writer110794014#yago:AmericanPeopleOfItalianDescent#yago:TwinPeople#ya
go:ItalianImmigrantsToTheUnitedStates#yago:AmericanActorsOfSwedishDescen
t#yago:ItalianPeopleOfGermanDescent#yago:Person100007846#yago:SwedishPeo
ple#yago:FilmMaker110088390#yago:Philanthropist110421956#yago:ItalianPeo
pleOfSwedishDescent#yago:ItalianFemaleModels#http://umbel.org/umbel/rc/A
ctor#yago:Model110324560yago:ItalianFilmActors"
    }
  ]
}
```


GetImage

Per ogni risorsa contenuta nel JSON, la *webapp* lancia una chiamata al servizio getImage passando come parametri i valori di "uri" e di "title". GetImage ritorna un JSON di questo tipo:


```
[
 {
  "imageURL":
"http://upload.wikimedia.org/wikipedia/en/0/09/Roberto_Rossellini.jpg"
 }
]
```


L'immagine presente all'indirizzo specificato, viene passata allo *step* di visualizzazione, insieme alla stringa "Label" contenuta nell'output di Classify (è il nome della risorsa che verrà visualizzato sull'immagine). Una volta ottenuti i risultati di tutte le chiamate a GetImage, la *webapp* chiama in sequenza per ogni risorsa contenuta nel JSON di Classify i seguenti servizi di TMF Server: getText, getNews, getVideo (se "mergedType" contiene *DBpedia:Actor*, *DBpedia:Activity*,

DBpedia:Band, *DBpedia:Artist* tranne *DBpedia:Writer*, *DBpedia:Athlete*, *DBpedia:MusicalWork*, *Freebase:/celebrities* o *Freebase:/film/actor*) e infine getMap (se "mergedType" contiene *DBepdia:Place*).

GetText

L'output di GetText è un JSON con la seguente struttura:

```
[
 {
  "title": "Roberto Rossellini"
 }
]
```

Il valore di "title" viene inserito in una chiamata alle API di Wikipedia per ottenere il testo della pagina di Wikipedia a cui appartiene quel titolo, insieme al valore della lingua scelta ("it" o "en"):

http://<lingua>.wikipedia.org/w/api.php?redirects&callback=?&action=parse&page=<title>&format=json

Tale chiamata restituisce un JSON contenente il testo preformattato dell'articolo di Wikipedia in questione, che viene passato allo *step* di visualizzazione.

GetNews

L'output di getNews è il seguente:

```
{
     "results":[
      {
         "body":"The most important line in President Obama's Afghan
speech was not about Afpak policy (so named by the White House) but
about the U.S. domestic situation: "Our troop commitment in Afghanistan
cannot be open-ended - because the nation that I am most interested in
building is our own." As military strategy for winning a war",
         "date":"20091208",
         "des_facet":[
            "UNITED STATES DEFENSE AND MILITARY FORCES",
            "UNITED STATES ECONOMY",
            "AFGHANISTAN WAR (2001- )"
         ],
```

```
         "title":"OP-ED COLUMNIST; Afghanistan on Main Street",
         "url":"http:\/\/www.nytimes.com\/2009\/12\/08\/opinion\/08iht-
edcohen.html"
      },
      {
         "body":"The top military commander in Afghanistan warns in a
confidential assessment of the war there that he needs additional troops
within the next year or else the conflict ''will likely result in
failure.'' The grim assessment is contained in a 66-page report that the
commander, Gen. Stanley A. McChrystal, submitted to Defense Secretary
Robert M. Gates",
         "date":"20090921",
         "des_facet":[
            "UNITED STATES DEFENSE AND MILITARY FORCES",
            "AFGHANISTAN WAR (2001- )"
         ],
         "title":"General Calls for More Troops To Avoid Afghanistan
Failure",

"url":"http:\/\/www.nytimes.com\/2009\/09\/21\/world\/asia\/21afghan.htm
l"
      }
   ]
}
```


I valori di "body", "title" e "url" vengono estratti e passati allo *step* di visualizzazione.

GetMap

L'output di getMap è il seguente:


```
[
 {
  "lat": "41.900002",
  "long" : "12.483334"
 }
]
```


I valori "lat" e "long" vengono utilizzati per creare un oggetto Map della libreria Javascript Leaflet[194] centrato sul punto desiderato e "zoomato" a una certa

[194] URL: http://leaflet.cloudmade.com/

distanza. A questo oggetto viene aggiunto un *layer* di OpenStreetMap[195] per aumentare le informazione spaziali disponibili. Questo oggetto viene passato allo *step* di visualizzazione.

GetVideo

L'output di getVideo è il seguente:

```
[
 {
  "videoURL": "http://youtu.be/Ytf4oyPSb1Y"
 }
]
```

Il valore di "videoURL" viene passato allo *step* di visualizzazione.

GetMetadata

Dopo i servizi di TMF server, la *webapp* chiama il metodo interno getMetadata per ottenere i metadati messi a disposizione da DBpedia per la risorsa in questione. Questo metodo controlla se l'URI della risorsa si riferisce a DBpedia o DBpedia Italia: nel primo caso le query sono lanciate sullo SPARQL endpoint di DBpedia[196] nel secondo caso sullo SPARQL endpoint di DBpedia Italia[197]. Per ogni risorsa le query sono due, una per le *object properties* e una per le *datatype properties* (al posto di <risorsa> viene passato il valore di "uri" dell'output di Classify).

Object properties[198]:

```
select ?subjectLabel ?relationLabel ?objectLabel where {
<risorsa> ?relation ?object .
<risorsa> <http://www.w3.org/2000/01/rdf-schema#label> ?subjectLabel .
?relation <http://www.w3.org/2000/01/rdf-schema#label> ?relationLabel .
?object <http://www.w3.org/2000/01/rdf-schema#label> ?objectLabel .
FILTER (langMatches(lang(?subjectLabel), "EN")) .
```

[195] URL: http://www.openstreetmap.org/

[196] URL: http://dbpedia.org/sparql

[197] URL: http://it.dbpedia.org/sparql

[198] Nel caso le query siano lanciate sull'endpoint di DBpedia Italia, la stringa "EN" è sostituita con "IT".

```
FILTER (langMatches(lang(?relationLabel), "EN")) .
FILTER (langMatches(lang(?objectLabel), "EN")) .
}
```

Datatype properties:

```
select ?subjectLabel ?relationLabel ?object where {
<risorsa> ?relation ?object .
<risorsa> <http://www.w3.org/2000/01/rdf-schema#label> ?subjectLabel .
?relation <http://www.w3.org/2000/01/rdf-schema#label> ?relationLabel .
?relation <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2002/07/owl#DatatypeProperty>
FILTER (langMatches(lang(?subjectLabel), "EN")) .
FILTER (langMatches(lang(?relationLabel), "EN")) .
}
```

La *webapp* ha una *blacklist* di *object properties* e *datatype properties* da scartare perché di scarso interesse per l'utente nel contesto della visualizzazione dei metadati (es: http://dbpedia.org/ontology/abstract, http://www.w3.org/2000/01/rdf-schema#comment, ecc.). Le triple risultanti dalle due query viste in precedenza sono mergiate in un'unica lista e filtrare sulla base della *blacklist*. Le relazioni e gli oggetti delle triple rimanenti sono quindi passate allo *step* di visualizzazione.

3.7.4.3 Visualizzazione dei contenuti

Per la visualizzazione dei contenuti viene utilizzata la libreria per JQuery FancyBox[199], uno strumento che offre un modo elegante per aggiungere funzionalità di zoom per le immagini o i contenuti HTML e multimediali alle pagine Web. La FancyBox viene attivata al click dell'utente su di un frame della griglia degli argomenti. Man mano che la *webapp* ottiene i contenuti (o i link ai contenuti) dai servizi di TMF server, questi vengono embeddati nella FancyBox in maniera diversa a seconda del contenuto, come descritto di seguito.

L'immagine viene passata alla FancyBox come URL. Il box si adatta per rappresentare l'immagine, trovando un compromesso tra la dimensione dello schermo e quella dell'immagine stessa.

[199] URL: http://fancybox.net/

1) Il testo di Wikipedia viene passato come HTML + CSS, con un particolare carattere e sfondo che simula l'effetto di un libro cartaceo.
2) Il testo del New York Times viene passato come HTML + CSS, con un particolare carattere e sfondo che simula l'effetto di un quotidiano cartaceo.
3) La mappa viene passata come oggetto Map della libreria Leaflet per generare la visualizzazione del punto desiderato all'interno di un riquadro di dimensioni fisse.
4) Il video viene passato come *snippet* HTML che incorpora un SWF di YouTube.
5) Per la visualizzazione dei metadati viene effettuato l'embedding all'interno della FancyBox della pagina HTML in cui viene costruita la struttura grafica. La libreria JavaScript utilizzata per creare tale struttura è InfoVis Toolkit[200] che mette a disposizione un insieme di strumenti per creare visualizzazioni di dati interattivi per il Web.

[200] URL: http://thejit.org/

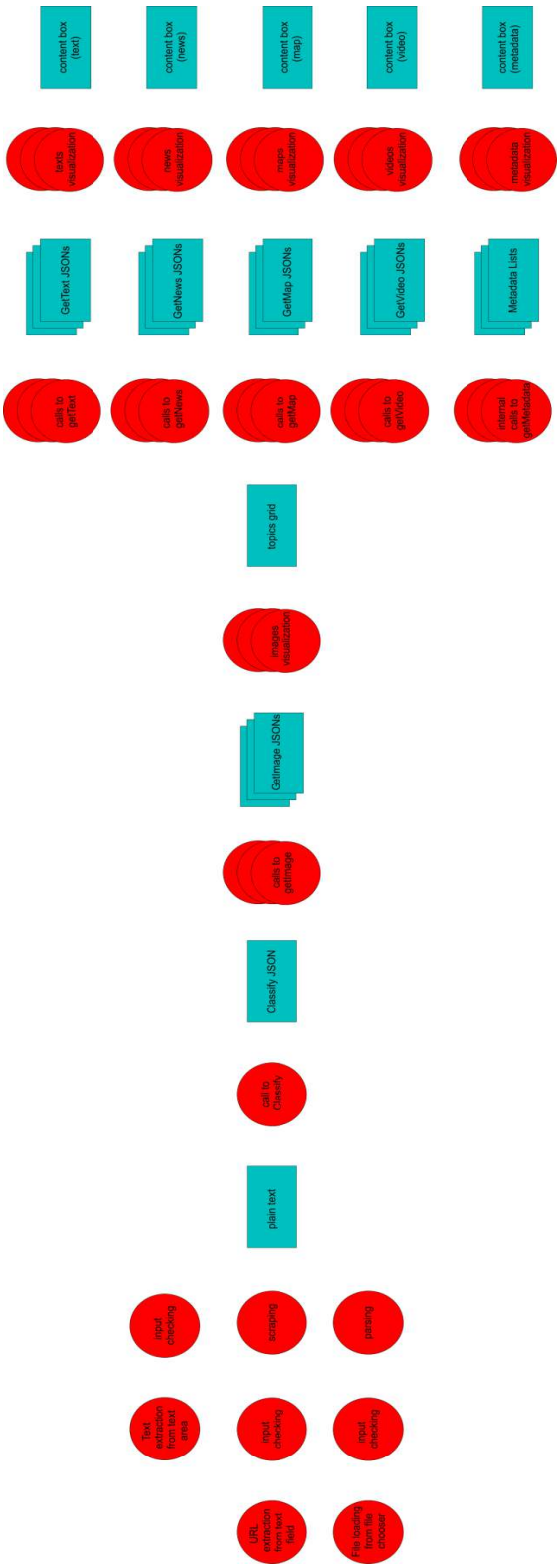


Figura 32 - Schema sintetico del flusso di esecuzione di TMF

3.8 Test e prestazioni

Per testare le performance di TMF, è stato utilizzato un corpus di 50 testi online in italiano e 50 testi online in inglese di lunghezza compresa tra le 100 e le 5000 parole (principalmente articoli di quotidiano e di riviste specialistiche). Il metodo utilizzato per l'acquisizione del testo è stato per tutti lo *scraping* da URL[201]. Per ogni lingua, i testi sono stati sottoposti a TMF prima in sequenza e poi in parallelo in batterie da 5, 10 e 30. Le misurazioni effettuate sono state:

1) tempo impiegato per un singolo documento dal momento del *submit* al momento in cui la GUI visualizza l'ultima immagine nella griglia degli argomenti[202] (Tabella 9 e 10, colonna 3)
2) tempo impiegato per un singolo documento dal servizio Classify (tabella 9 e 10, colonna 4)
3) tempo impiegato per un documento (inserito in una batteria di 5 documenti in parallelo) dal momento del submit al momento in cui la GUI visualizza l'ultima immagine nella griglia degli argomenti (Tabella 9 e 10, colonna 5)
4) tempo impiegato per un documento (inserito in una batteria di 5 documenti in parallelo) dal servizio Classify (Tabella 9 e 10, colonna 6)
5) tempo impiegato per un documento (inserito in una batteria di 10 documenti in parallelo) dal momento del submit al momento in cui la GUI visualizza l'ultima immagine nella griglia degli argomenti (Tabella 9 e 10, colonna 7)
6) tempo impiegato per un documento (inserito in una batteria di 10 documenti in parallelo) dal servizio Classify (Tabella 9 e 10, colonna 8)

[201] Il metodo di acquisizione (digitazione diretta nella *text area*, inserimento dell'URL nella *text field* o caricamento di un file) si è osservato non modificare in misura rilevante le performance del sistema a parità di lunghezza del testo. Lo *scraping* da URL e il *parsing* di file contenenti fino a 5000 parole sono eseguiti dal TMF Server in pochi millesimi di secondo.

[202] È stato scelto questo intervallo di tempo perché coincide col blocco della navigazione dell'utente sulla GUI: dal *submit* del *form* fino al caricamento dell'ultima immagine, la GUI non consente all'utente di continuare la navigazione. La griglia tuttavia comincia a comparire (con un effetto di dissolvenza in entrata) non appena la prima immagine viene caricata, dunque il tempo di attesa visivo dell'utente è minore del tempo di attesa relativo all'interazione. Dopo lo sblocco della griglia, ci sono tempi di attesa minimi per l'utente, in quanto tutti gli altri contenuti vengono caricati in *background* a partire dalla fine del caricamento dell'ultima immagine.

7) tempo impiegato per un documento (inserito in una batteria di 30 documenti in parallelo) dal momento del submit al momento in cui la GUI visualizza l'ultima immagine nella griglia degli argomenti (Tabella 9 e 10, colonna 9)
8) tempo impiegato per un documento (inserito in una batteria di 30 documenti in parallelo) dal servizio Classify (Tabella 9 e 10, colonna 10)

La macchina su cui sono stati effettuati i test è una workstation con CPU quad-core da 3,2 GHz e RAM 12 GB, su cui gira il sistema operativo Ubuntu 11.10. La memoria riservata alla JVM è di 10 GB, tuttavia il processo di TMF server non ha mai utilizzato più di 6 GB di RAM durante la sua esecuzione. I test si sono svolti all'interno della LAN del Centro Nexa su Internet e Società del Politecnico di Torino, dunque la connessione non ha prodotto tempi aggiuntivi.

Di seguito la tabella riassuntiva dei risultati medi (secondi/parole) per i documenti italiani e inglesi, con relativo diagramma a barre. In Appendice vengono invece riportate tutte le misure effettuate nei test, con relativi diagrammi a linee.

	Media tempo totale per parola	**Media tempo Classify per parola**	**Media tempo totale per parola (batteria 5 doc)**	**Media tempo Classify per parola (batteria 5 doc)**	**Media tempo totale per parola (batteria 10 doc)**	**Media tempo Classify per parola (batteria 10 doc)**	**Media tempo totale per parola (batteria 30 doc)**	**Media tempo Classify per parola (batteria 30 doc)**
Italiano	0,011 s/w	0,004 s/w	0,038 s/w	0,020 s/w	0,049 s/w	0,021 s/w	0,097 s/w	0,024 s/w
Inglese	0,012 s/w	0,005 s/w	0,037 s/w	0,022 s/w	0,051 s/w	0,023 s/w	0,099 s/w	0,024 s/w

Tabella 8 - Medie tempi per parola (italiano e inglese)

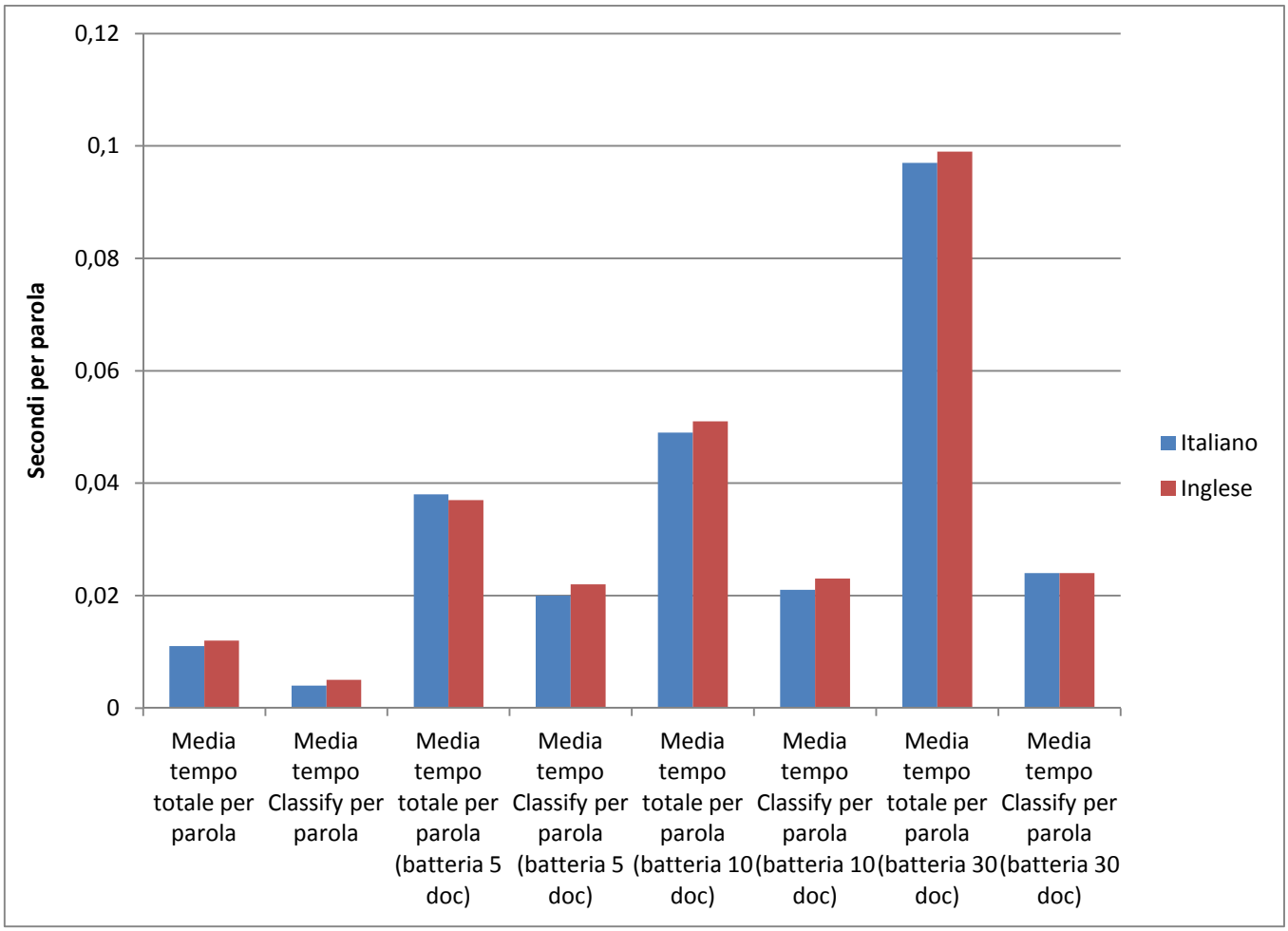


Figura 33 - Diagramma Medie per paROLA (Italiano e Inglese)

I risultati hanno evidenziato una sostanziale somiglianza nei tempi per i testi in lingua italiana e inglese.

Un testo di 500 parole (in italiano o inglese) viene classificato in circa 4 secondi ed arricchito in circa 7 secondi[203]. Le prestazioni relative (secondi/parole) migliorano all'aumentare del numero di parole del documento. Un testo di 5000 parole viene classificato in circa 8 secondi ed arricchito in circa 11 secondi. La media secondi/parole è di circa 0.005 per la classificazione e di 0.011 per l'intero processo.

[203] Qui e di seguito si intende per tempo di arricchimento il tempo complessivo di classificazione + *enhancement*.

Quando i testi vengono inviati al sistema in parallelo, in batterie da 5 documenti, si assiste a un aumento dei tempi sia della classificazione sia dell'arricchimento. Un testo di 500 parole inviato contemporaneamente ad altri 4 documenti viene classificato in circa 9 secondi ed arricchito in circa 18 secondi. Anche in questo caso le prestazioni relative (secondi/parole) migliorano all'aumentare del numero di parole del documento. Un testo di 5000 parole inviato contemporaneamente ad altri 4 documenti viene classificato in circa 12 secondi ed arricchito in circa 21 secondi. La media secondi/parole è di circa 0.020 per la classificazione e di 0.038 per l'intero processo.

Quando i testi vengono inviati al sistema in parallelo, in batterie da 10 documenti, si assiste a un ulteriore aumento dei tempi dell'arricchimento, mentre i tempi di classificazione rimangono pressoché invariati. Un testo di 500 parole inviato contemporaneamente ad altri 9 documenti viene classificato in circa 10 secondi ed arricchito in circa 22 secondi. Anche in questo caso le prestazioni relative (secondi/parole) migliorano all'aumentare del numero di parole del documento. Un testo di 5000 parole inviato contemporaneamente ad altri 9 documenti viene classificato in circa 12 secondi ed arricchito in circa 24 secondi. La media secondi/parole è di circa 0.021 per la classificazione e di 0.049 per l'intero processo.

Quando i testi vengono inviati al sistema in parallelo, in batterie da 30 documenti, si assiste a un forte aumento dei tempi dell'arricchimento, ma ad un aumento molto lieve dei tempi di classificazione. Un testo di 500 parole inviato contemporaneamente ad altri 29 documenti viene classificato in circa 11 secondi ed arricchito in circa 40 secondi. Anche in questo caso le prestazioni relative (secondi/parole) migliorano all'aumentare del numero di parole del documento. Un testo di 5000 parole inviato contemporaneamente ad altri 29 documenti viene classificato in circa 12 secondi ed arricchito in circa 47 secondi. La media secondi/parole è di circa 0.024 per la classificazione e di 0.097 per l'intero processo.

Il sistema mostra quindi una buona scalabilità relativa alla classificazione, ma non altrettanto buona in relazione all'arricchimento. La causa potrebbe essere una parametrizzazione inefficace degli oggetti Lucene Reader e Lucene Searcher nei servizi di arricchimento (getImage, getText, getNews ecc), che non concorro-

no correttamente in parallelo: si cercherà di porre rimedio a questa issue nei rilasci successivi del software.

Documento	**Num di parole**	**Tempo totale singolo**	**Tempo Classify singolo**	**Tempo totale in batteria da 5 doc**	**Tempo Classify In batteria da 5 doc**	**Tempo totale in batteria da 10 doc**	**Tempo Classify in batteria da 10 doc**	**Tempo totale in batteria da 30 doc**	**Tempo Classify In batteria da 30 doc**
1	433	4,2	3,1	15,9	9,1	19,4	8,1	38,9	10,3
2	644	5,3	3,2	16,0	10,2	21,8	9,8	50,8	12,1
3	2902	8,1	7,8	20,7	10,5	27,1	11,6	45,2	13,5
4	4779	10,6	8,1	22,9	12,6	31,8	12,7	52,9	14,3
5	567	5,1	3,4	17,2	9,7	20,5	9,6	40,8	11,6
6	723	6,2	4,5	19,8	9,0	24,6	10,0	46,4	12,0
7	525	8,2	4,6	22,1	9,8	25,9	10,8	49,3	13,8
8	230	4,0	2,7	15,8	9,0	15,5	6,9	37,3	9,0
9	369	4,7	2,5	15,9	9,1	16,8	6,0	35,3	8,5
10	448	5,6	3,9	18,0	9,7	18,9	7,9	38,6	10,1
11	611	6,3	3,1	18,5	10,1	27,4	9,1	63,5	11,0
12	605	6,1	3,1	17,9	9,8	27,5	9,4	50,6	12,7
13	994	7,8	5,4	23,8	10,2	26,0	11,0	49,3	13,2
14	350	4,4	2,2	16,1	9,2	16,4	5,6	34,3	8,5
15	1266	8,4	6,1	22,0	10,5	29,4	11,3	44,5	13,5
16	3047	9,7	7,8	20,9	11,6	30,8	13,0	50,0	12,7
17	560	10,9	4,0	19,8	10,1	19,1	15,2	40,8	11,7
18	149	10,0	2,1	14,8	8,9	18,9	6,9	40,3	8,8
19	990	7,7	5,2	23,5	9,8	25,0	10,6	49,4	13,0
20	459	10,7	3,5	14,5	10,4	18,0	5,1	33,9	8,0
21	1376	8,7	5,9	16,8	9,3	26,4	11,9	50,8	12,8
22	682	6,4	4,0	16,0	10,0	25,7	9,8	43,6	11,9
23	210	4,3	2,9	13,6	8,6	19,5	9,9	35,7	8,0
24	651	8,9	3,0	15,6	9,0	18,6	8,3	61,2	12,7
25	1022	8,4	5,1	22,7	10,2	26,4	11,0	44,6	13,5
26	242	4,4	3,1	14,6	9,5	20,9	12,4	34,3	13,1
27	685	9,1	3,4	15,4	9,3	18,7	8,8	60,8	13,9
28	770	6,9	4,0	17,7	9,9	19,0	12,0	54,3	12,7
29	509	6,7	4,0	20,5	9,4	17,0	11,8	50,6	12,2

30	597	7,7	3,9	19,3	8,7	17,2	11,0	48,1	11,1

Tabella 9 - Risultati per i documenti in italiano

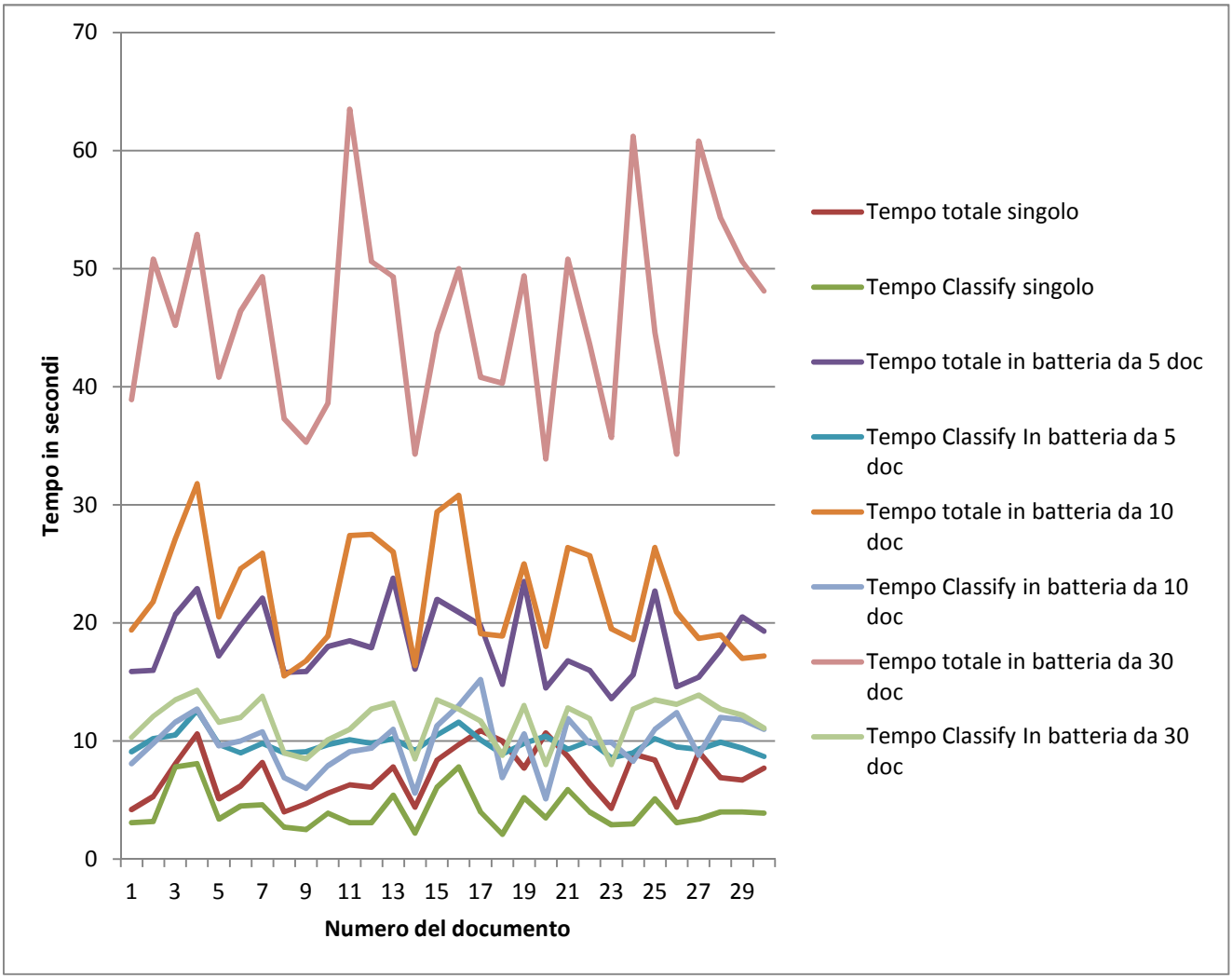


Figura 34 - Diagramma a Linee risultati italiano

Documento	Num di parole	Tempo totale singolo	Tempo Classify singolo	Tempo totale in batteria da 5 doc	Tempo Classify In batteria da 5 doc	Tempo totale in batteria da 10 doc	Tempo Classify in batteria da 10 doc	Tempo totale in batteria da 30 doc	Tempo Classify In batteria da 30 doc
1	576	5,5	3,6	17,0	9,6	20,8	9,2	40,8	11,6
2	711	6,0	4,1	19,6	9,0	24,6	10,7	46,1	12,1
3	521	8,2	4,6	23,1	9,8	21,9	10,8	50,1	13,7
4	232	4,2	2,9	15,7	9,0	15,2	7,0	40,0	9,1
5	433	4,2	3,1	15,9	9,1	19,4	8,1	38,9	10,3
6	644	5,3	3,2	16,0	10,2	21,8	9,8	50,8	12,1
7	2784	9,4	6,1	20,7	10,5	27,1	11,6	45,2	13,5
8	4502	11,2	8,5	23,9	12,8	31,9	11,7	51,4	15,4
9	378	4,4	2,2	16,9	9,5	17,9	6,8	34,9	8,0
10	450	5,5	3,8	18,7	9,5	18,0	7,0	35,7	10,3
11	611	6,3	3,1	18,5	10,1	27,4	9,1	63,5	11,0
12	605	6,1	3,1	17,9	9,8	27,5	9,4	50,6	12,7
13	984	7,3	5,2	22,7	11,3	27,5	12,0	49,9	14,5
14	156	10,3	2,4	13,8	8,5	17,5	7,5	41,3	8,5
15	921	7,7	5,3	24,5	9,6	25,9	11,6	50,4	13,5
16	460	10,7	3,5	14,5	10,4	18,0	5,1	33,9	8,0
17	1465	8,9	6,0	17,8	9,6	28,4	12,4	52,8	13,8
18	360	4,4	2,2	15,7	9,2	18,4	5,8	32,3	8,5
19	1270	8,6	6,2	22,5	10,6	28,4	12,3	41,5	13,7
20	3159	9,8	7,9	23,9	11,7	32,8	13,7	54,0	12,8
21	566	10,9	4,0	19,8	10,1	19,1	15,2	40,8	11,7
22	676	6,4	4,3	16,4	11,0	24,7	9,8	47,6	11,9
23	214	4,3	2,9	13,6	8,6	19,5	9,9	35,7	8,0
24	657	8,9	3,0	15,6	9,0	18,6	8,3	61,2	12,7
25	765	6,9	4,0	17,7	9,9	19,0	12,0	54,3	12,7
26	523	6,7	4,0	20,5	9,4	17,0	11,8	50,6	12,2
27	601	7,7	3,9	19,3	8,7	17,2	11,0	48,1	11,1
28	1098	8,5	5,2	23,7	10,8	25,4	11,2	42,6	13,5
29	245	4,4	3,1	14,6	9,5	20,9	12,4	34,3	13,1
30	693	9,1	3,4	16,4	9,3	18,9	8,8	66,8	13,9

Tabella 10 - Risultati per i documenti in inglese

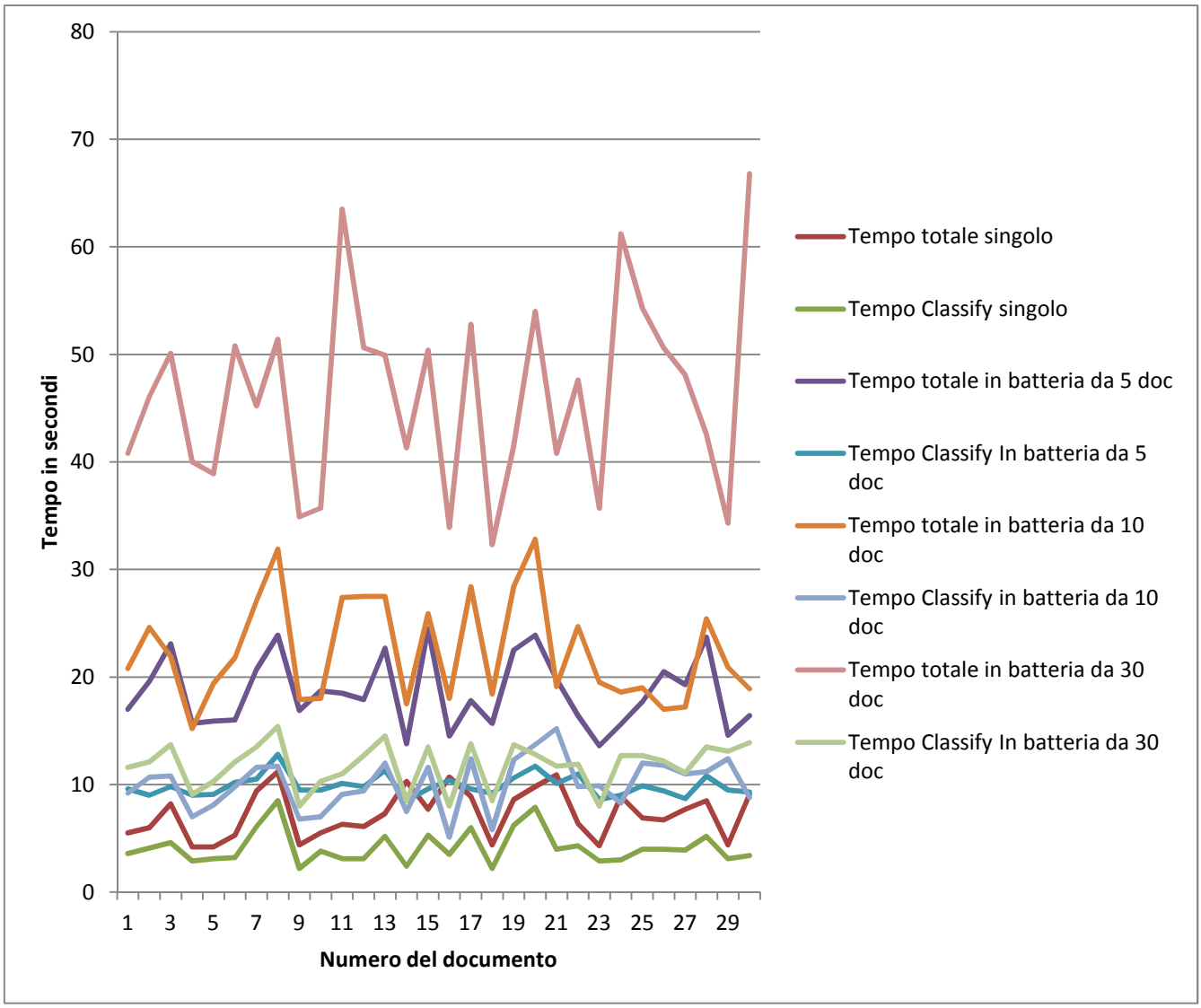


Figura 35 - Diagramma a linee risultati inglese

Capitolo 4

DBpedia Gateways contro l'information overload sul Web

4.1 Il problema dell'information overload in Rete

L'espressione *information overload*, tradotta in italiano con "sovraccarico cognitivo" o "sovraccarico informativo", è stata utilizzata per la prima volta dallo scrittore e giornalista americano Alvin Toffler nel saggio *Future Shock* (1970). Ovviamente il concetto non aveva nulla a che vedere con Internet, ma si riferiva ai disagi psico-comportamentali cui poteva andare incontro il cittadino della grande metropoli, bombardato da un flusso continuo di informazioni mediatiche, di *stressor* lavorativi, di ansie postmoderne.

> Managers plagued by demands for rapid, incessant and complex decisions; pupils deluged with facts and hit with repeated tests; housewives confronted with squalling children, jangling telephones, broken washing machines, the wail of rock and roll from the teenager's living room and the whine of the television set in the parlor—may well find their ability to think and act clearly impaired by the waves of information crashing into their senses. (Toffler, 1970, p. 181)

Benché il contesto sia diverso, la definizione di *information overload* data da Toffler può oggi risultare adatta a descrivere un problema tipico dell'età dell'informazione, ovvero la difficoltà da parte dell'utente dei media digitali (in particolar modo di Internet) di utilizzare la massiccia mole di informazioni con cui viene a contatto quotidianamente. Per Toffler esiste un limite nella «quantità di informazione» che il cervello umano più assorbire, manipolare, valutare e trattenere oltre il quale le *performance* cognitive dell'uomo precipitano inesorabil-

mente. Considerando la mente umana come un «canale» che riceve degli input (le informazioni), le elabora e produce degli output (le decisioni, le azioni), esperimenti hanno dimostrato che all'aumento eccessivo degli input corrisponde un peggioramento degli output in termini di precisione ed efficacia.

> Information has been defined technically and measured in terms of units called "bits." By now, experiments have established rates for the processing involved in a wide variety of tasks from reading, typing, and playing the piano to manipulating dials or doing mental arithmetic. And while researchers differ as to the exact figures, they strongly agree on two basic principles: first, that man has limited capacity; and second, that overloading the system leads to serious breakdown of performance. (Toffler, 1970, p. 181)

L'*information overload* si ripercuote dunque sulla capacità dell'uomo di prendere decisioni in base alle proprie conoscenze. La nozione di sovraccarico cognitivo indagata da Toffler insiste particolarmente sull'aspetto sincronico: gli stimoli cognitivi che avvengono nello stesso momento o in un intervallo di tempo molto breve hanno effetti peggiori, a parità di numero, rispetto a quelli che si presentano in modo distanziato nel tempo. Conferma di questa tesi proviene dagli studi nell'ambito della neurobiologia sulla capacità del cervello umano di reagire agli stimoli. Nel saggio *The Overflowing Brain: Information Overload and the Limits of Working Memory* (2009) il neuroscienziato svedese Torkel Klingberg esplora i limiti del cervello umano nel trattenere l'informazione nella memoria operativa (*working memory*) all'aumentare degli input provenienti dal mondo esterno. La quantità e la complessità dell'informazione da gestire nella vita quotidiana è accresciuta in maniera esponenziale nel mondo digitale, ma il cervello dell'uomo contemporaneo è anatomicamente e chimicamente lo stesso dei primi Homo Sapiens. La tecnologia ci permette di accedere più facilmente, in ogni luogo e in ogni momento, a dati e informazioni di ogni tipo, ma i ritmi di apprendimento, la capacità di concentrarsi e di ricordare, non sono cambiate sostanzialmente da 200.000 anni a questa parte, e dunque il cervello non riesce a sfruttare efficacemente tutto questo materiale informativo.

> The brains with which we are born today are almost identical to those with which Cro-Magnons were born forty thousand years ago. If there is some inherent limitation to our ability to handle information, it should be present already at this time, when the most technologically advanced artifact was the barbed bone harpoon. The same brain now has to take on the torrent of information that the digital society discharges over us. A Cro-Magnon met in one year as many people as you and I can meet in one day. The volume and complexity of the information we're expected to handle continues to increase. (Klingberg, 2009, p. 10-11)

Il meccanismo che il cervello attiva quando si sente sovraccaricato è quello di "rilasciare" le informazioni ritenute inessenziali: vuoti di memoria temporanei, errori, distrazioni più o meno gravi sono il risultato di questa reazione. Ma nel lungo periodo lo stress cognitivo (chiamato *infostress* da Klingberg) può dare origine a problemi medici come tutte le altre forme di stress. Per quanto diversi studi abbiano dimostrato che il livello medio del quoziente intellettivo sia in costante crescita dal 1930 ad oggi (Flynn, 1987 e 1999), da cui si può probabilmente dedurre una certa adattabilità del cervello all'aumento costante della quantità di stimoli, molte altre ricerche testimoniano che le malattie da stress sono diventate il male endemico delle popolazioni occidentali (Sapolsky, 1994). Nel celebre *Why Zebras Don't Get Ulcers*, il biologo americano Robert Sapolsky spiega che il rischio maggiore quando l'uomo è messo di fronte a situazioni che eccedono le proprie capacità di gestirle è quello dell'impotenza appresa («learned helplessness»), un possibile prodromo della depressione. Questo meccanismo si attiva quando si è sottoposti per un lungo periodo di tempo a uno *stressor* impossibile da fronteggiare, tanto da abituarsi all'impotenza e smettere di cercare una soluzione ai problemi. L'incapacità di gestire il flusso di informazioni nelle professioni a più alto carico decisionale è certamente fonte di stress prolungato e può portare nei casi più gravi a forme di impotenza appresa.

> It takes surprisingly little in terms of uncontrollable unpleasantness to make humans give up and become helpless in a generalized way. In one study by Donald Hiroto, student volunteers were exposed to either escapable or inescapable loud noises (as in all such studies, the two groups were paired so that they were exposed to the same amount of noise). Afterward, they were given a learning task in which a correct

> response turned off a loud noise; the "inescapable" group was significantly less capable of learning the task. (Sapolsky, 1994, p. 155)

Il rapporto tra lo stress e le sfide cognitive affrontate è sicuramente condizionato delle proprie capacità e competenze. Lo psicologo ungherese Mihaly Csikszentmihalyi (1990) ha analizzato il concetto di *flow* (flusso), che è la sensazione che proviamo quando siamo completamente focalizzati e assorbiti nel lavoro che stiamo svolgendo. Per esempio un pianista assorto nell'esecuzione di un nuovo pezzo, che diventa inconsapevole di se stesso e del trascorrere del tempo, è in uno stato di *flow*. Questa specie di "trance agonistica" può essere raggiunta anche quando un giocatore di scacchi risponde mossa su mossa facendo uso di tutte le proprie capacità ed esperienze o da un medico nel mezzo di un'operazione chirurgica. Quello che Csikszentmihalyi prova a fare è identificare le circostanze che conducono al *flow*. Egli dimostra che se analizziamo le situazioni lavorative e competitive nei termini delle sfide che presentano e delle capacità delle persone impegnate ad affrontarle, scopriamo che il *flow* si origina in contesti caratterizzati da un alto livello di sfida e di capacità, nei quali le competenze di chi agisce combaciano perfettamente con le esigenze del compito che viene eseguito. Considerando il diagramma di Csikszentmihalyi (Figura 36) come una mappa cognitiva con il nord nel quadrante superiore, è nel settore di nord-est che troviamo lo stato di *flow*. Quando le sfide eccedono le competenze, abbiamo l'ansia e lo stress. Quando al contrario le competenze eccedono le sfide, abbiamo il senso di controllo che si trasforma gradatamente in noia al diminuire delle sfide. Basta scambiare la parola "competenza" con "capacità di memoria operativa" e la parola "sfida" con "carico informativo" per ottenere una mappa della percezione soggettiva dell'*information overload* (Klingberg, 2009). Quando lo sforzo richiesto per elaborare e trattenere le informazioni supera le nostre capacità, facciamo esperienza dei problemi di concentrazione e di memoria posizionati nella zona nord-ovest della mappa. È solo nei casi in cui le sfide cognitive e la nostra abilità di affrontarle sono in uno stato di equilibrio tra loro che può realizzarsi lo stato di *flow*, in cui apprendiamo nuova conoscenza senza sovraffaticare la nostra mente e con perfetta naturalezza.

In our studies, we found that every flow activity, whether it involved competition, chance, or any other dimension of experience, had this in common: It provided a sense of discovery, a creative feeling of transporting the person into a new reality. It pushed the person to higher levels of performance, and led to previously undreamed-of states of consciousness. In short, it transformed the self by making it more complex. In this growth of the self lies the key to flow activities. (Csikszentmihalyi, 1990, p. 74)

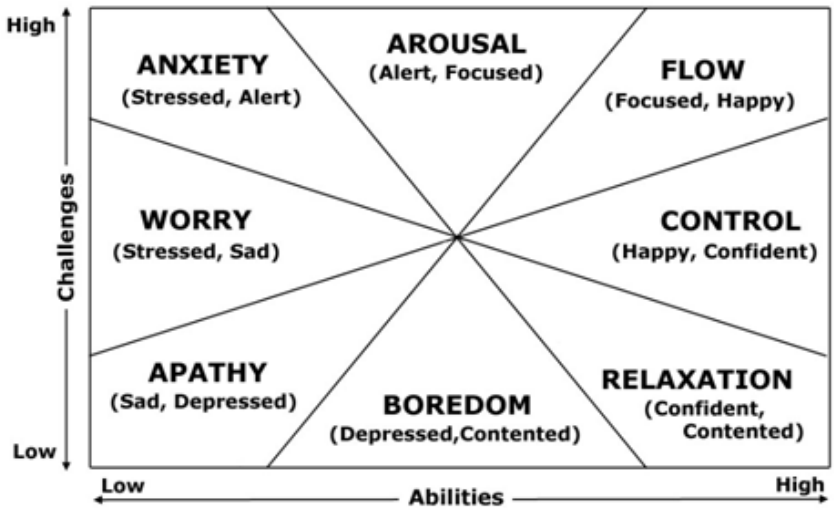


Figura 36 - Diagramma challenges-abilities (Csikszentmihalyi, 1990, p. 74)

Emerge in maniera piuttosto chiara che il problema dell'information overload in era digitale è connesso alla capacità di filtrare le informazioni, in quanto il flusso di dati potenzialmente disponibile in Rete è infinitamente maggiore rispetto all'abilità del cervello umano di farvi fronte. Come ha espresso Clay Shirky in una conferenza al Web 2.0 Expo 2008[204], «It's not information overload. It's filter failure». La selezione dell'informazione è un'attività che affonda le sue radici in epoca pre-moderna. Nel testo *Too Much to Know: Managing Scholarly Information before the Modern Age* (2010) la storica Ann Blair fa notare che nell'antichità e nel Medioevo vi erano già studiosi che lamentavano la sovrabbondanza di libri a disposizione ed esistevano strumenti per l'*information management* come raccolte di sommari, bibliografie, estratti dai testi, ecc, organizzati per argomento e meticolosamente conservati. Quando la diffusione della

[204] URL: http://youtu.be/LabqeJEOQyI

stampa nella metà del XV Secolo ha accresciuto massicciamente il numero e la disponibilità di libri, una fascia sempre più ampia di persone istruite ha sentito la necessità di trovare rimedio ad una situazione percepita come sovraccarico di conoscenza. Tra il Cinque e il Seicento si è quindi assistito alla nascita dei primi *reference book* (opere di consultazione), antenati delle moderne enciclopedie.

> The authors of reference books present themselves as compilers, responsible for the accurate reporting of what others had written elsewhere but not for the veracity of those statements themselves. Compilers were therefore conveyors of information rather than of their own opinions or positions. […] Early compilations involved various combinations of four crucial operations: storing, sorting, selecting, and summarizing, which I think of as the four S's of text management. We too store, sort, select, and summarize information, but now we rely not only on human memory, manuscript and print as in earlier centuries. (Blair, 2010, p. 2-3)

Questi strumenti servivano per filtrare l'informazione più rilevante (secondo un criterio umano) e scartare quella ritenuta superflua, consentendo al lettore di concentrare i propri sforzi su un flusso cognitivo meno ricco. Per usare i termini di Csikszentmihalyi, "alleggerivano la sfida" in maniera che le "competenze" bastassero ad affrontarla senza produrre uno stato di stress. I *reference book* non erano utili soltanto perché sintetizzavano in (relativamente) poche frasi vasti ambiti di conoscenza, ma perché consigliavano, spesso in rigoroso ordine di importanza, una serie di testi da cui attingere per approfondire o controllare le informazioni riferite. È questo il genere di selezione di cui ha bisogno il lettore per poter formare la propria cultura e il proprio gusto. La stessa selezione è operata quotidianamente dalle biblioteche, sia specialistiche che generiche, per acquisire nuovi libri o liberarsi di altri. L'*output* di nuovo materiale pubblicato, in tutte le forme, è talmente cospicuo che nessuna libreria, nemmeno la più grande, può sperare di acquisirlo tutto; anche in settori relativamente specializzati un qualche tipo di filtro è necessario e la maggior parte delle librerie ha precise politiche di selezione (Kujoth, 1969). Diverse indagini sono state condotte sui modi in cui gli esperti bibliotecari decidono quali nuovi testi acquisire ogni anno sulla base della percezione che hanno dell'evoluzione dei vari campi di studio, e si è scoperto la modalità più frequente è la discussione informale tra i colleghi (Kovacs, 1990). È

quindi sulla base dell'esperienza, del *know-how* di esperti di dominio, che la conoscenza viene filtrata fino all'utente finale, che si ritrova sgravato da un compito che senza questo processo avrebbe dovuto compiere autonomamente.

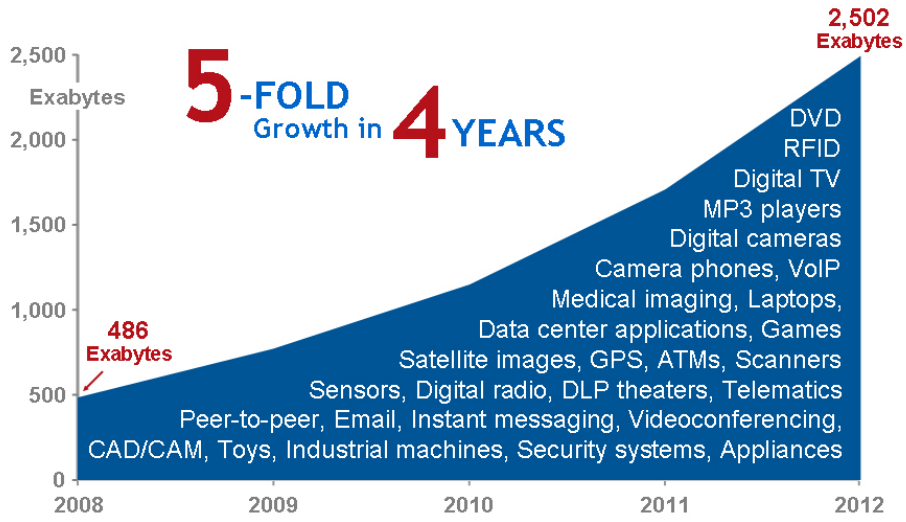


Figura 37 - crescita dell'informazione digitale prodotta nel mondo dal 2008 al 2012[205]

Nello spazio informativo del Web, i Linked Open Data possono contribuire efficacemente a contrastare il problema del sovraccarico informativo spesso lamentato dagli utenti. È soprattutto questo l'aspetto del Web Semantico che viene indagato da David Weinberger nel capitolo 9 di *Too Big To Know* (2011). Il testo di Weinberger ha come argomento principale la conoscenza umana e il modo in cui si è modificata nel passaggio dai sistemi editoriali classici (l'editoria cartacea in prima battuta, ma anche la radio e la televisione, in quanto caratterizzati dallo stesso principio restrittivo di filtraggio dei contenuti) all'universo del World Wide Web. Ciò che interessa all'autore non è stabilire se l'infrastruttura di Internet stia peggiorando o migliorando il nostro modo di conoscere il mondo, perché questa valutazione implicherebbe un concetto univoco di conoscenza impossibile da raggiungere, ma tentare di comprendere in che modo questa infrastruttura possa essere migliorata per esaltare i pregi del mezzo Internet, e limitarne i difetti, in relazione all'acquisizione di conoscenza.

[205] URL: http://www.emc.com/images/about/news/press/2009/createreplicate-large.jpg

Le caratteristiche che contraddistinguono in Web rispetto agli altri media sono per Weinberger essenzialmente cinque:

1) Abbondanza di informazione. Sul Web è disponibile una quantità di informazione incomparabile rispetto ai media precedenti come la televisione e la carta stampata. Basti pensare al numero di testi digitalizzati dal progetto Google Books (più di 15 milioni) o al numero medio di documenti ritornati cercando su un motore di ricerca in rete le parole di uso comune (mediamente nell'ordine delle centinaia di milioni). Si stima che la quantità di informazione digitale prodotta sul pianeta sia cresciuta dal 2008 ad oggi da 486 exabyte (1 exabyte = 10^18 byte) a circa 2.500 exabyte[206] (vedi Figura 37).
2) Ricchezza di collegamenti. Nel corso dei secoli la conoscenza, fa notare Weinberger, è sempre stata interconnessa attraverso citazioni, note, rimandi, bibliografie, ecc. Il Web ha portato all'estremo questa caratteristica facendo del link il centro della fruizione dell'intero medium: è il link che da continuità alla nostra esperienza sul Web, che rende il Web un luogo unitario e non solo un insieme di risorse separate tra loro.
3) Libertà di espressione. Weinberger sottolinea che lo spazio informativo del Web è *permission-free*: non occorre un'autorizzazione per pubblicarci sopra qualsiasi cosa, ovviamente nei limiti della legalità. Non esiste, come in passato, un sistema autoritativo che impedisce ad alcune risorse di essere diffuse (la redazione di un giornale, il comitato di una casa editrice, la produzione di un'emittente televisiva, ecc), facendo da filtro tra i creatori dei contenuti e il pubblico.
4) Pubblico accesso. A meno di non voler apportare particolari restrizioni ai propri contenuti, essi sono "pubblici" di *default* sul Web, ricercabili e fruibili da ogni altro utente. Spesso non soltanto il prodotto finito della nostra comunicazione è reso pubblico, ma anche il processo attraverso il quale viene prodotta. Si pensi alle pagine "Talk" e "View history" di ogni voce di Wikipedia, ai *pre-print* dei paper, degli articoli scientifici o delle linee guida: possiamo seguire passo passo il loro evolversi nel tempo, dalla prima bozza alla versione finale.
5) Conoscenza irrisolta. Il tipo di conoscenza presente sul Web non è organica, sistematica, finalizzata a un obiettivo ultimo. Da questo punto di

[206] URL: http://www.emc.com/images/about/news/press/2009/createreplicate-large.jpg

> vista il Web è specchio di un relativismo in cui può coesistere tutto e il contrario di tutto e sopratutto ogni cosa è sullo stesso piano. Le pagine Web hanno la stessa identica struttura formale (codice HTML, un po' di *scripting* e qualche abbellimento grafico) qualsiasi argomento trattino (popolare o accademico, sacro o profano) e la loro distanza dall'utente è quantificabile nello stesso numero di click. È anche vero che ormai, scrive ironicamente Weinberger, «We have lived through enough fundamental revolutions of thought to suspect that we don't happen to be living in the age that finally gets everything right» (Weinberger, 2011, p. 181).

Queste caratteristiche fanno emergere come fondamentale tendenza del Web l'"inclusività", laddove i media tradizionali, a cominciare dalla stampa, favoriscono l'"esclusività". Nel mondo della carta stampata l'esclusione di risorse informative comincia da ciò che gli editori decidono di stampare, per ragioni principalmente economiche: lo sforzo economico dietro alla pubblicazione di un volume cartaceo infatti è considerevole e deve essere ripagato dai guadagni della sua vendita. Gli editori filtrano le proposte di pubblicazione più promettenti scartando quelle che pensano non possano garantire un guadagno adeguato, così come le librerie e le biblioteche pubbliche espongono nello spazio limitato dei loro scaffali i volumi che possono interessare maggiormente i propri utenti. La televisione e la radio adottano lo stesso meccanismo di filtro per riempire lo spazio dei palinsesti, perché soggette alle regole dell'*audience* e della concorrenza. Internet invece non presenta alcun problema materiale per la pubblicazione di una risorsa informativa, se non quello di possedere un computer e un accesso alla Rete. Per questo la dimensione del Web è aumentata nel tempo a un ritmo frenetico, passando da circa 26 milioni di pagine indicizzate nel 1998 a circa 1000 miliardi di pagine nel 2008[207] (Alpert et al., 2008), appartenenti a più di 110 milioni di domini diversi (vedi Figura 38). Il lato oscuro dell'abbondanza di informazione sul Web è noto con il nome di *information overload*, ovvero l'incapacità di comprendere ed utilizzare efficacemente l'informazione a causa della sua eccessiva quantità. L'*information overload* si lega a un altro problema generato dalla libertà di espressione e dalla disorganicità del Web: quello dell'*authority*, ovvero

[207] Stiamo parlando di una crescita del 38 mila per cento.

l'autorevolezza e affidabilità delle informazioni. Nei sistemi editoriali tradizionali l'autorevolezza è garantita sia dall'autore stesso (dalla sua carica o dalla fama come esperto di qualche argomento) sia dall'editore. Un articolo apparso su una prestigiosa rivista scientifica o un romanzo edito in una collana di grandi successi spesso costituisce una garanzia per il pubblico anche indipendentemente dal nome dell'autore. Essere pubblicati da un certo editore è ancora oggi per alcuni scrittori, ricercatori, artisti, il traguardo fondamentale dell'intera carriera. Pubblicare un documento su Internet invece non dice nulla di per sé sulla qualità del documento. È necessario rendere esplicita, attraverso metadati, l'informazione sull'autorevolezza che in precedenza era garantita dal mezzo.

> The difference between the sentence "Birds descended from dinosaurs" when it comes from some anonymous stranger on the Internet and when it comes from *Nature* is the metadata that says *Nature* is reliable. It used to be that the authority metadata was implicit in how the knowledge was distributed: it came from *Nature*, or from your physician's mouth. The mere existence of a book from a reputable publisher was metadata that established that at least some authorities thought it was worthwhile. Since simply being posted in a permission-free world conveys zero metadata about authority, that metadata now has to be made far more explicit. (Weinberger, 2011, p. 179)

La maggior parte dell'informazione presente online oggi non ha un'attribuzione efficace della fonte. Per attribuire la fonte a una risorsa sul Web sono necessari tre elementi:

1) un identificatore univoco di una fonte di informazione, ovvero un puntatore che si riferisce in maniera non ambigua a un produttore di informazione, come centro di ricerca, un'istituzione, un'azienda o un singolo individuo;
2) una relazione di attribuzione che colleghi la risorsa informativa all'identificativo della fonte;
3) una certificazione della fonte, ovvero un'ulteriore relazione che colleghi l'attribuzione della fonte all'identificativo di un ente certificatore (che può essere la fonte stessa o un ente terzo).

L'architettura odierna del Web rende difficile il procedimento di attribuzione della fonte, prima di tutto per l'assenza di identificativi univoci di oggetti reali, che sono invece la base del Web Semantico. Nella maggior parte dei casi la fonte è attribuita in maniera implicita (identificata con il *copyright* del sito Web che pubblica la risorsa) o attraverso semplici stringhe (nei commenti o nei forum) o ancora per mezzo di link al "profilo" di un soggetto o di un ente (*username* utilizzati per una specifica piattaforma, siti istituzionali, ecc.).

La sfida per una nuova infrastruttura della conoscenza su Web è riuscire a mantenere intatti i tratti fondamentali di Internet, che ne hanno determinato il suo successo a livello planetario, limitando il più possibile gli svantaggi che il medium inevitabilmente porta con sé. I Linked Data sono indubbiamente uno strumento utile in questa direzione.

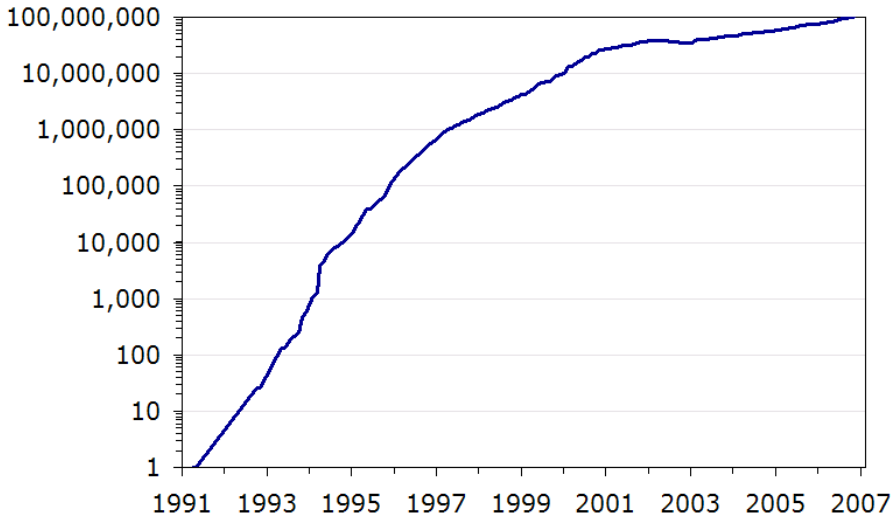


Figura 38 - Crescita del numero di domini sul Web dal 1991 al 2007[208]

Se facciamo una ricerca su Google e ci vengono proposti 10 milioni di risultati, non siamo di per sé di fronte a un problema di *information overload*, ma lo siamo se Google non ci fornisce insieme ai risultati gli strumenti necessari per comprendere quali di essi maggiormente contano per noi o quali di essi siano più affidabili in relazione all'autorevolezza della fonte.

[208] URL: http://www.nngroup.com/articles/100-million-websites/

La strategia per sconfiggere il sovraccarico informativo, sostiene Weinberger, è paradossalmente quella di aggiungere informazione all'informazione: «The solution to the information overload problem is to create more information: metadata». Il Semantic Web viene interpretato da Weinberger soprattutto come tentativo di metadatare i contenuti del Web attraverso concetti non ambigui rappresentati da URI. Se gli argomenti trattati in una pagina Web sono identificati univocamente attraverso gli URI, anziché per mezzo di semplici tag, è più facile per un motore di ricerca rispondere alle nostre esigenze, riducendo i sinonimi a uno stesso concetto o eliminando l'ambiguità di alcune parole. Ma ancora più importanti sono i metadati che riguardano la fonte di un documento sul Web, in quanto rispondono all'esigenza di ricostituire in parte quel sistema di filtraggio dell'informazione che permetteva ai sistemi dell'editoria classica di funzionare così bene.

> Metadata helps with the second problem inherent in an open superabundant system: most of what's posted will be crap. So, we need ways to evaluate and filter which can be especially difficult since what is crap for one effort may be gold for another. [...] Indeed, a little metadata can go a very long way. This is important because in the Net of abundance we need more metadata about the authority of works that credentialed institution can provide. (Weinberger, 2011, p. 185)

Adottando l'approccio dei Linked Data, soggetti giuridici come le aziende, le fondazioni, gli enti di ricerca, le istituzioni, ma anche i singoli individui, diventano URI a cui è possibile riferirsi in maniera non ambigua. Essi sono collegati da appositi predicati alle triple RDF che esprimono una qualche informazione in maniera da costituirne la fonte. Questo procedimento è chiamato *reification* e ne possiamo vedere un esempio in Figura 39. L'informazione che Ora Lassila è autore della pagina Web http://www.w3.org/Home/Lassila è data dalla tripla:

```
<Ora Lassila> <http://description.org/schema/Creator>
<http://www.w3.org/Home/Lassila> .
```

Quest'affermazione viene reificato in un blank node[209] (l'ovale vuoto al centro), diventando un oggetto di tipo Statement che ha tre particolari proprietà RDF: *rdf:subject* (il soggetto della frase, cioè <Ora Lassila>), *rdf:predicate* (il predicato della frase, cioè <http://description.org/schema/Creator>) e *rdf:object* (il complemento oggetto della frase, ovvero <http://www.w3.org/Home/Lassila>). A queste tre proprietà se ne aggiunge una quarta che è appunto l'attribuzione della fonte: <a:attributedTo> <Ralph Swick>. Ralph Swick è dunque la fonte secondo la quale Ora Lassila sarebbe autore della pagina Web http://www.w3.org/Home/Lassila.

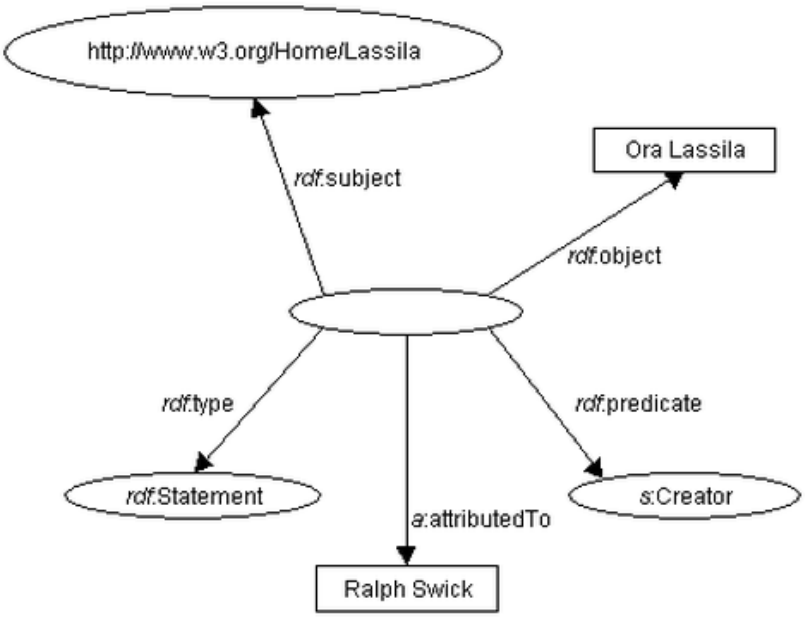


Figura 39 - Reification e attribuzione della fonte in RDF[210]

Facilitare, come fa il Web Semantico, l'inserimento di informazioni aggiuntive riguardanti il significato (la semantica appunto) dei documenti su Internet e la loro fonte (attribuita in maniera non ambigua attraverso URI nel dominio di enti certificati), costituisce sicuramente un passaggio importante per migliorare la nuova infrastruttura della conoscenza veicolata dal Web.

[209] URL: http://www.w3.org/TR/rdf-concepts/#dfn-blank-node
[210] URL: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/#higherorder

4.2 DBpedia Gateways

Per DBpedia Gateways si intende qui una forma di organizzazione della conoscenza online basata sulla classificazione *DBpedia-driven* delle risorse Web. Limitando, a titolo esemplificativo, il numero dei documenti presenti sul Web ai contenuti di lingua inglese, si considerano idealmente come documenti target per la classificazione tutte le risorse contenenti testo in lingua inglese accessibili in Rete (HTML, PDF, DOC, ecc.), e come profili delle classi tutte le voci della versione inglese di Wikipedia. La KB che fornisce le categorie della classificazione semantica è dunque DBpedia, per cui ogni risorsa di DBpedia (corrispondente a una voce di Wikipedia) rappresenta il profilo di una classe foglia, mentre le classi padre sono rappresentate dai nodi dell'ontologia di DBpedia[211] da cui le altre classi ereditano. Il risultato del processo di classificazione, virtualmente infinito nel tempo, è l'assegnazione di ogni documento in lingua inglese presente sul Web a un certo numero (prestabilito in fase di *thresholding*) di categorie semantiche identificate attraverso URI di DBpedia. Queste categorie dovrebbero indicare il significato o i principali argomenti del testo target.

Il processo di categorizzazione può avvenire sia in forma manuale che automatizzata. Nel caso di procedimento manuale, si tratta di assegnare a una risorsa testuale sul Web un certo numero di argomenti sotto forma di voci di Wikipedia. Essendo noti il contenuto della risorsa da catalogare e il contenuto della voce candidata di Wikipedia, il compito del classificatore umano è quello di stabilire, sulla base di un'interpretazione personale, se il significato del testo è affine a quello della voce, ovvero se l'argomento trattato nella voce può essere considerato come argomento anche del testo target. Il classificatore stila una lista di voci di Wikipedia ordinata per rilevanza rispetto al testo: la corrispondenza tra le voci di Wikipedia e gli URI di DBpedia è facilmente derivata dal sistema attraverso il dataset *Links to Wikipedia Article* rilasciato dal progetto DBpedia[212].

Se il processo è automatizzato, invece, occorre scegliere un modello di classificazione automatica (si vedano gli approcci visti al paragrafo 2.3.1) e implemen-

[211] L'ontologia di DBpedia è navigabile online all'indirizzo: http://mappings.dbpedia.org/server/ontology/classes/

[212] URL: http://wiki.dbpedia.org/Downloads38#links-to-wikipedia-article

tarlo via software. TellMeFirst implementa il modello k-Nearest Neighbor attraverso il framework Lucene, ma è possibile trovare soluzioni altrettanto o ancor più precise ed efficienti che implementino algoritmi diversi, come Support Vector Machine o Latent Semantic Analysis. Negli approcci basati sul Machine Learning viene considerato come *training set* l'insieme delle voci di Wikipedia, dato che ogni articolo di Wikipedia è quanto meno pre-etichettato per mezzo del suo titolo (ulteriori annotazioni possono essere considerate i *wikilink*, come in TellMeFirst e in DBpedia Spotlight).

Scegliendo TellMeFirst come classificatore e 5 come soglia RCut, un breve documento come quello seguente sarebbe categorizzato con gli URI:

1) http://dbpedia.org/resource/Renewable_energy (2,456)
2) http://dbpedia.org/resource/Carbon_capture_and_storage (1,897)
3) http://dbpedia.org/resource/Efficient_energy_use (1,1776)
4) http://dbpedia.org/resource/Greenhouse_gas (1,750)
5) http://dbpedia.org/resource/Nuclear_power (1,345)

> Industry letter calls for decarbonisation target in energy bill (The Guardian, 05/11/2012)[213]
>
> Leaders of renewables, nuclear and CCS groups warn government is putting new jobs and financial investment at risk. New jobs and financial investment in the energy sector are at risk if the government does not ensure its imminent energy bill supports low-carbon power, an unusual coalition of the trade bodies representing the renewable energy, nuclear power and carbon capture industries has warned. The letter to the energy and climate secretary, Ed Davey, strongly backs the call from the government's climate advisers, the Committee on Climate Change (CCC), to include a reference in the bill for the power sector to be almost entirely decarbonised by 2030. On Saturday, the Observer revealed that the amount of power expected to be generated from gas by 2030 has quadrupled in the last year, raising fears that carbon targets will be missed and low-carbon generation crowded out. The final version of the energy bill, due to bring in the biggest reforms to the energy market in two decades, is expected towards the end of November. The draft version was published in May. The letter signed by the heads of RenewableUK, the Carbon Capture and Storage Association and the Nuclear Industry Association says that a decarbonisation reference would lower "the perceived

[213] URL: http://www.guardian.co.uk/environment/2012/nov/05/letter-decarbonisation-target-energy-bill

political risks, but could also reduce the cost of capital for decarbonising the power sector. We therefore believe that this could be very important for investment going forward." The CCC has recommended electricity in 2030 be produced at no more than 50g of CO2/kW by 2030; gas power stations emit around 350g of CO2/kW. The current energy bill draft has no such target. The trade body chiefs, who represent more than 1,000 companies between them, say that any significant delay in the bill "could result in investment being postponed, with major implications for associated new industrial development and jobs in a high-tech, high growth sector." John Sauven, the executive director at Greenpeace, which opposes nuclear power, took the surprising step of welcoming the letter, saying: "This letter shows that whilst different industries will have differing preferences for the exact mix of energy technologies, there is unity from across huge swathes of the business community on the need for a clear goal in the energy bill to take carbon almost completely out of the electricity system by 2030."

Al contrario della classificazione manuale, quella automatica produce di fianco alle categorie risultanti uno *score*, ovvero un valore numerico che indica il grado di similarità del documento target col profilo della classe candidata. Questo *score* è interessante ai fini dell'organizzazione della conoscenza sul Web, in quanto rende esplicito il grado di sicurezza (*confidence*) di una singola categorizzazione per un documento. Una volta uniformati i criteri di *scoring* tra i vari sistemi di classificazione, sarà possibile integrarne i risultati sulla base di un coefficiente che potremmo chiamare "coefficiente di *confidence*" della classificazione. Mettiamo il caso che un secondo software di classificazione, basato su un algoritmo diverso da k-NN, dia come risultato per il testo visto in precedenza questa lista:

1) http://dbpedia.org/resource/Carbon_capture_and_storage (2,679)
2) http://dbpedia.org/resource/Geothermal_energy (1,1776)
3) http://dbpedia.org/page/Greenhouse_gas (1,750)
4) http://dbpedia.org/resource/Renewable_energy (1,234)
5) http://dbpedia.org/page/Solar_energy (1,163)

Si tratta di integrare le due liste in modo da produrre una terza che tenga conto dei risultati di entrambi i classificatori. Gli URI ripetuti fanno media tra loro, ottenendo come output:

1) http://dbpedia.org/resource/Carbon_capture_and_storage (2,288)
2) http://dbpedia.org/resource/Renewable_energy (1,845)
3) http://dbpedia.org/resource/Efficient_energy_use (1,1776)
4) http://dbpedia.org/resource/Geothermal_energy (1,654)
5) http://dbpedia.org/page/Greenhouse_gas (1,750)
6) http://dbpedia.org/page/Solar_energy (1,163)
7) http://dbpedia.org/resource/Nuclear_power (1,345)

Replicando questo processo per ogni classificazione automatica risultante da un diverso classificatore si ottiene una categorizzazione unificata via software di quel documento sul Web. Tuttavia bisogna tenere conto anche delle classificazioni manuali che ogni documento può subire nel corso del tempo, una fonte di informazione estremamente ricca, spontanea e in molti casi affidabile. L'obiettivo è quello di trasformare le liste di classificazione generate dagli utenti in *ranked lists* ordinate secondo lo stesso modello di *scoring* dei sistemi automatici di classificazione, in modo da poter integrare la classificazione *user-generated* con quella automatica. Utilizzando una tecnica ben nota ai progettisti dei siti di *social tagging* (o *social bookmarking*, come Digg[214], Delicious[215], Reddit[216], ecc.), il *rank* di una classificazione, anche se non specificato dall'utente, può essere inferito in maniera statistica sulla base della classificazione che quella stessa risorsa ha subito da parte degli altri utenti. Il coefficiente di *confidence* di una certa categorizzazione è, nell'ambito del *social tagging*, funzione del numero di utenti che hanno scelto quella stessa categoria rispetto al numero totale di utenti che hanno categorizzato il contenuto (Trant, 2009; Milicevic et al., 2010; Gupta et al., 2010). L'integrazione col sistema di *scoring* dei classificatori automatici può avvenire normalizzando i risultati *human-generated* e quelli *machine-generated* in modo che esprimano valori reali compresi tra 0 e 1, poi combinandoli secondo un criterio che può privilegiare l'affidabilità della classificazione umana o quella automatica. Ogniqualvolta una nuova classificazione, umana o automatica, si aggiunge a una risorsa sul Web in forma di URI di DBpedia, ne viene calcolato lo *scoring* e viene aggiunta alla risorsa in questione.

[214] URL: http://digg.com/
[215] URL: https://delicious.com/
[216] URL: http://www.reddit.com/

Il risultato finale dell'intero procedimento può essere visto come un insieme di documenti sul Web metadatati con categorie di Wikipedia/DBpedia, ma anche come una serie di *hub* che, a partire da specifici argomenti di Wikipedia/DBpedia, conducono a *repositories* di risorse informative riguardanti quegli argomenti. Le possibilità offerte da questo modello di organizzazione della conoscenza sono molteplici. In primo luogo, a partire da qualsiasi voce di Wikipedia, si accede una vasta mole di documenti ordinati secondo la rilevanza rispetto a quell'argomento. Ogni articolo di Wikipedia può contenere dunque il link a un motore di ricerca semantico specifico per un argomento, dove la ricerca può essere successivamente affinata per autore, data, fonte, dimensione e tipo di documento, ecc. I depositi specialistici di contenuti sono appunto i "DBpedia Gateways", che costituiscono il punto di collegamento tra una risorsa di DBpedia/Wikipedia, semanticamente univoca e identificato attraverso un URI, e lo sterminato insieme di risorse informative sul Web ad essa riconducibili.

Il criterio di ricerca navigazionale (detto *category browsing* o *faceted browsing*) può affiancarsi come valido concorrente del modello *query-based* utilizzato per esempio da Google. Essendo le categorie della classificazione organizzate secondo un'ontologia, sarà l'esplorazione dell'ontologia a condurre l'utente verso l'oggetto che maggiormente soddisfa il proprio bisogno informativo. Strumenti grafici per rendere semplice e intuitiva la navigazione di un'ontologia sono di supporto a questo processo: esempi sviluppati di recente sono BubbleTree[217] di Open Knowledge Foundation, LodLive[218] e ICONVIS[219] del Politecnico di Torino. In particolar modo ICONVIS (Cairo et al., 2011), prevede tutti gli *step* del processo di Information Retrieval, dall'esplorazione delle classi del grafo (le categorie padri) alla visualizzazione delle loro istanze (le classi foglia) e delle loro relazioni (i predicati presenti nell'ontologia), fino all'accesso diretto ai documenti precedentemente classificati sotto le rispettive istanze in un database locale.

> ICONVIS consente all'utente di visualizzare con estrema semplicità tutti i dati delle istanze di una specifica classe o di una singola istanza a partire dal grafo ontologico.

[217] URL: http://okfnlabs.org/bubbletree/
[218] URL: http://en.lodlive.it/
[219] URL: http://iconvis.polito.it/

Esplorando i concetti di un'ontologia sui beni culturali, ad esempio, è possibile raggiungere la classe "Palazzo". Nel caso in cui l'utente sia interessato ai dati del DB delle istanze della classe 'Palazzo', può visualizzarle tramite un "interruttore" contrassegnato da un'icona bianca, riconducibile alla forma di un DB. Il grafo tassonomico si riduce in alto a sinistra poco prima che i dati vengano visualizzati sullo schermo, in modo tale che l'informazione relativa all'esplorazione dei concetti da parte dell'utente non venga persa. (Futia, 2011, p. 75)

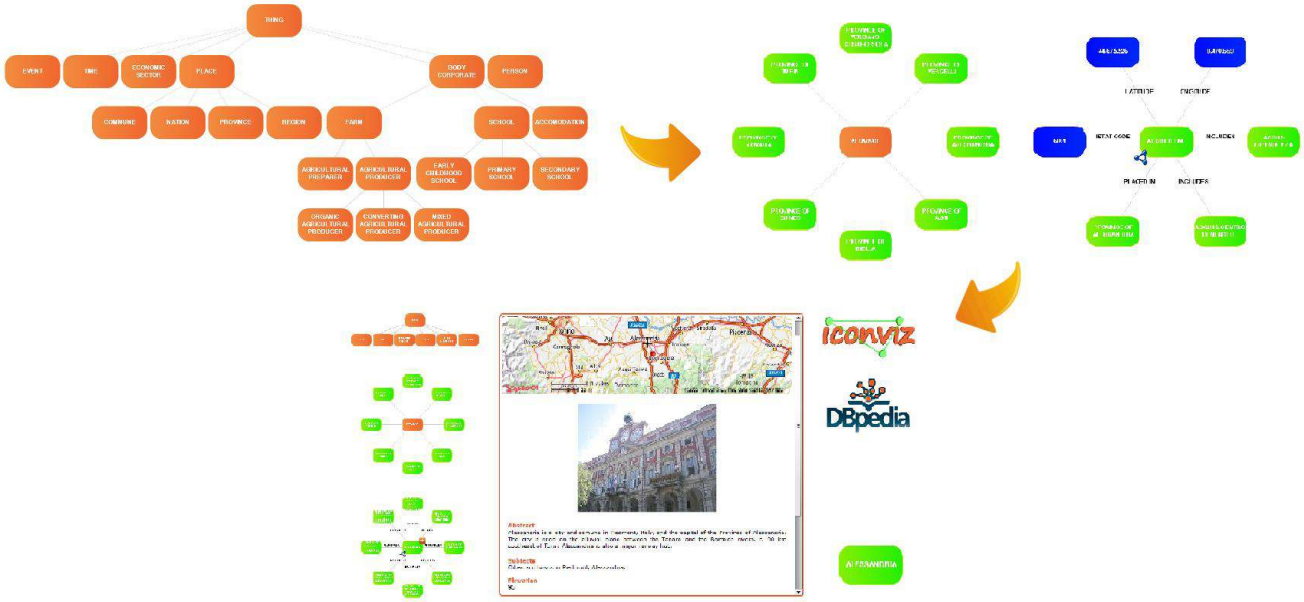


Figura 40 – Visualizzazione di un'ontologia e dei dati in essa classificati attraverso ICONVIS

Un'alternativa alla navigazione grafica dello schema ontologico è la ricerca per filtri (o "a faccette"), che consente di visualizzare le categorie della classificazione in un elenco e di scegliere quali di esse includere o escludere dalla ricerca. Naturalmente, selezionando una classe padre, vengono automaticamente incluse le classi foglia, da cui si possono poi deselezionare gli elementi indesiderati. Considerata la ricchezza della KB di DBpedia, questo metodo consente una ricerca per argomenti estremamente più precisa della ricerca per *keyword* tradizionale. Esempio recente di un sistema di motore di ricerca *ontology-based* dove prevale la ricerca a faccette è OPSA DOC[220]. OPSA è un progetto europeo nato nel 2012

[220] URL: http://www.opsa.eu/cms/it/opsadoc.html

con lo scopo di favorire la condivisione di conoscenze fra gli attori locali e le popolazioni delle regioni italo-francesi Piemonte, Liguria, Côte d'Azur e Rhône-Alpes attraverso la creazione di una rete transfrontaliera d'informazione e di scambi in materia di sanità pubblica (osservazione sanitaria, pianificazione e promozione della salute). Uno dei risultati di OPSA è stata l'implementazione di una piattaforma unitaria di accesso alle risorse informative che sfrutta la KB "OPSA Ontology" nel dominio della promozione della salute per aggiungere potenzialità semantiche a un motore di ricerca full-text. OPSA Ontology è una risorsa concettuale bilingue (italiano e francese) composta da circa 200 classi, 50 diversi predicati e più di 7000 entità (istanze). È stata progettata dal tesista con l'ausilio di esperti di dominio italiani e francesi per descrivere formalmente i principali argomenti presenti nell'ambito della promozione della salute, in modo che l'utente, a partire da una esplorazione dei concetti, possa filtrare gli argomenti di maggior interesse fino ad accedere ai documenti sottostanti con la consapevolezza della ricchezza semantica dell'archivio.

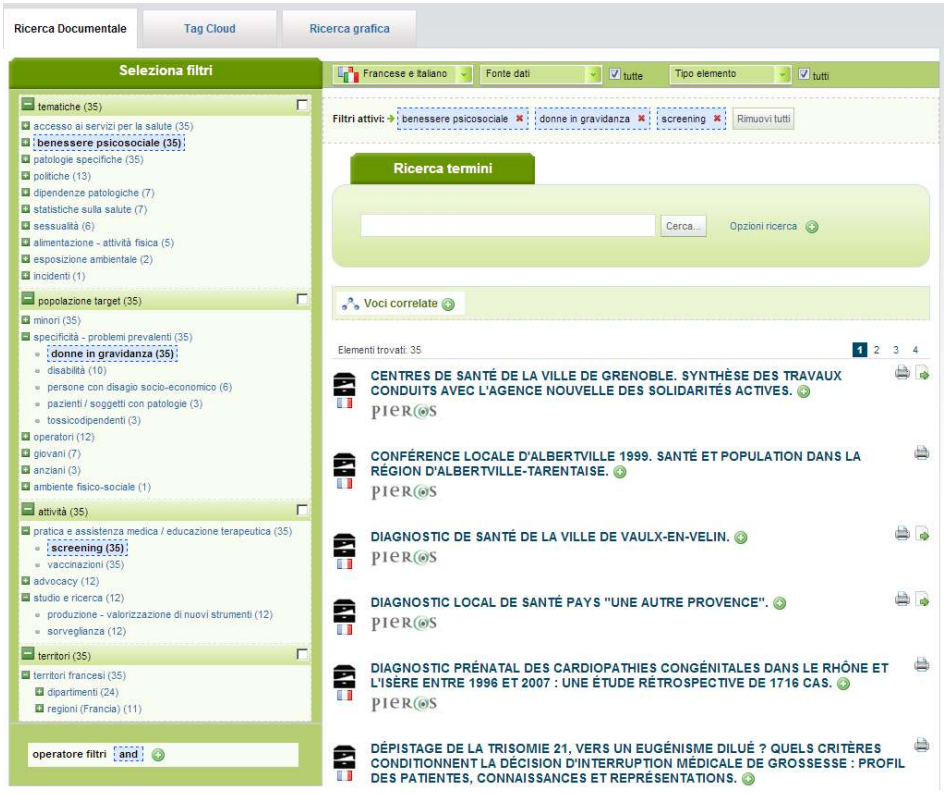


Figura 41 – Interfaccia del motore di ricerca ontology-based OPSA DOC

L'utilizzo di DBpedia/Wikipedia è il vero punto di forza dei DBpedia Gateways. DBpedia e Wikipedia sono strumenti in continua evoluzione[221]: ogni giorno nuove voci vengono aggiunte, modificate o eliminate, rispecchiando la costruzione collaborativa del sapere e la ricerca di un accordo semantico all'interno della comunità degli utenti. Allo stesso modo nuovi Gateways nasceranno ogni giorno, frutto di una conoscenza sempre più specialistica degli utenti della Rete, di nuovi accordi semantici e di nuove definizioni di significato (si veda l'esempio del "Future Internet Gateway" nel paragrafo 4.3). Il criterio della classificazione e del *ranking* deve essere sempre pubblico ed accessibile online, così come il modo in cui la classificazione da parte degli umani viene integrata con quella automatica. L'obiettivo dei DBpedia Gateway è infatti che l'organizzazione della conoscenza sul Web sia l'organizzazione che gli utenti stessi del Web danno a questa conoscenza. Al contrario dell'algoritmo di Google, gli approcci e le metodologie utilizzate per la costruzione dei Gateways non sono né opachi né imposti dall'alto. Se un utente non trova giusto vedere al primo posto in una categoria un certo documento, può assegnargli uno *score* più basso (*downvote*) e promuovere al primo posto un documento che gli sembra di maggior pertinenza o interesse. La comunità scientifica è sempre sollecitata a sviluppare classificatori automatici più precisi ed efficaci, rilasciando il codice *open source* in modo che chiunque possa contribuirvi. Tutte le informazioni riguardanti un certo documento possono essere aggiunte come metadati della risorsa informativa stessa, usando il semplice linguaggio RDF introdotto nel paragrafo 2.1. Tra queste informazioni: gli URI di classificazione con relativo *score*, il numero di classificazioni successive che il documento ha subito (e se sono umane o automatiche), i classificatori automatici sono stati utilizzati (sviluppati da chi e utilizzando quale algoritmo di *text classification*), ecc. Seguendo la prassi dei Linked Data, le informazioni aggiunte possono essere semplici valori *literal*, come ad esempio il numero di classificazioni, ma anche altri URI di DBpedia o di altri depositi di Linked Data, come il tipo di algoritmo di classificazione e l'ente di ricerca da cui è stato implementato.

[221] L'evoluzione di DBpedia segue ovviamente quella di Wikipedia. Le *release* di DBpedia hanno una cadenza periodica, ma esiste anche una versione "Live" della KB, che accoglie quasi in tempo reale ogni modifica operata su Wikipedia (Hellmann et al., 2009).

Il paragrafo successivo descrive la progettazione di un prototipo di DBpedia Gateways specifico per il campo della Future Internet Research.

4.3 Un progetto di esempio: Future Internet Gateway[222]

4.3.1 Introduzione

Future Internet Research and Experimentation (FIRE)[223] è un programma di ricerca finanziato dall'Unione Europea per promuovere gli sviluppi futuri delle tecnologie e dei servizi di Internet. Appartiene al settore ICT del settimo programma quadro (FP7), un programma di finanziamenti che copre il periodo 2007-2013 e rappresenta per l'Unione europea una opportunità di portare la sua politica di ricerca al livello delle sue ambizioni economiche e sociali.

La prima ondata di progetti FIRE è stata lanciata nel 2008, con un budget di 40 milioni di euro e con un forte focus sulle tematiche del *networking*. Nel 2010, una seconda generazione, con un budget di 50 milioni di euro, ha esteso in modo significativo il campo di applicazione verso tematiche riguardanti il Cloud Computing, le architetture orientate ai servizi, e le reti di sensori. La terza ondata ha preso il via nel 2011, con la creazione della rete di eccellenza EINS (Network of Excellence in InterNet Science[224]). Nel 2012, la call 8, con un bilancio di 25 milioni di Euro, ha aperto la strada ad una quarta generazione di progetti che sono iniziati in autunno. L'attenzione si è concentrata stavolta sulla condivisione di ambienti fisici e virtuali per la sperimentazione in campo Future Internet (FI) al fine di facilitare l'interscambio tra i progetti già esistenti.

---

[222] Il progetto Future Internet Gateway è nato con il nome di Internet Science Gateway in seguito alla conferenza "Verso una Internet Science" tenuta il 26/07/2012 a Trento da Juan Carlos De Martin e organizzata dalla Fondazione <ahref. La paternità dell'idea è da riconoscere a Juan Carlos De Martin, co-fondatore del Centro Nexa su Internet & Società, e a Luca De Biase, presidente di <ahref, con il contributo (in ordine alfabetico) di Federico Cairo, Stefano De Paoli, Raimondo Iemma, Michele Kettmaier, Federico Morando e Maurizio Teli. (URL: http://www.ahref.eu/it/events/a-cura-di-ahref/verso-una-internet-science)

[223] URL: http://cordis.europa.eu/fp7/ict/fire/

[224] URL: http://www.internet-science.eu/

La necessità più importante avvertita oggi in ambito FI è quella di riuscire a coordinare le iniziative e gli sforzi di diversi gruppi di ricerca, istituzioni e soggetti privati, per non disperdere le energie e per arrivare più velocemente ad obiettivi comuni. La FI è una linea di ricerca nuova, i cui confini non sono ancora stati definiti in maniera sistematica. È un settore estremamente articolato e multidisciplinare, in quanto vi convergono tematiche puramente tecniche, come le architetture di rete, il Cloud Computing, la *virtualization*, ecc, ricerche di tipo economico-giuridico, come la Internet Governance (Wilson, 2005), la Network Neutrality (Goth, 2010) e il diritto d'autore digitale (Stokes, 2009), e infine problematiche di carattere più sociologico, legate agli usi e alle funzioni sociali delle nuove tecnologie di Internet. I soggetti che trattano questi temi sono ovviamente molto eterogenei (centri di ricerca, soggetti istituzionali, aziende private, semplici studenti o appassionati) e hanno un *background* assai diversificato. La stessa cosa si può dire delle iniziative e della produzione documentale in quest'ambito: *paper* scientifici, conferenze, siti divulgativi, articoli di giornale, blog, ecc. Si tratta di materiale difficilmente riconducibile ad un'omogeneità concettuale, linguistica e di target.

4.3.2 I progetti di CSA in campo FIRE

Le iniziative di CSA (Coordination and Support Action) nell'ambito FIRE, finanziate a partire dal 2011, hanno mirato a costruire una prima rete di interconnessione tra i vari soggetti impegnati in tematiche FIRE. Si tratta di azioni che non riguardano la ricerca FI in sé, ma il coordinamento e la messa in rete di progetti, programmi e politiche relative a tale ricerca. I principali progetti di CSA per la tematica FIRE attualmente in corso sono AmpliFIRE[225], FUSION[226] e FIRE STATION[227]. Già conclusi sono invece MyFIRE[228], FIREBALL[229] e FI-

[225] URL: http://www.ict-fire.eu/home/amplifire.html
[226] URL: http://www.sme4fire.eu/
[227] URL: http://www.ict-fire.eu/home/firestation.html
[228] URL: http://www.my-fire.eu/
[229] URL: http://www.fireball4smartcities.eu/

REworks[230] (che è confluito in FIRE STATION). Di seguito sono sintetizzati gli obiettivi dei diversi progetti.

Lo scopo di MyFIRE è quello di favorire lo sviluppo di ambienti per la sperimentazione tecnologica in ambito FI in Europa e di diffondere *best practices* relative alle attività di test in questo settore. L'obiettivo vuole essere raggiunto attraverso la creazione di un dialogo aperto tra le comunità di ricercatori in tecnologie di rete e altri soggetti non tecnici come i responsabili politici e gli economisti.

FIREBALL istituisce un meccanismo di coordinamento attraverso il quale una rete di *smart cities* in Europa si impegna in un collaborazione di lunga durata per esplorare le opportunità della Internet del futuro. Il processo di coordinamento è supportato dall'utilizzo di una piattaforma sociale ("FIREBALL Community") che facilita il dialogo e la collaborazione tra le *smart cities* europee.

Il progetto AmpliFIRE intende effettuare una valutazione delle possibilità presenti e future della ricerca FIRE, individua le lacune esistenti e gli aspetti su cui è più importante evolvere. Sulla base di vari "Key Performance Indicators", AmpliFIRE monitora le condizioni tecniche, operative e organizzative dell'attuale ricerca FIRE nel suo complesso, per indicare i punti chiave su cui insistere per raggiungere gli obiettivi preposti dalla FI per il 2020.

L'obiettivo del progetto FUSION è invece quello di mettere in contatto le Piccole e Medie Imprese interessate ad utilizzare nuovi ambienti di collaudo con gli enti di ricerca o istituzionali che stanno sviluppando o hanno sviluppato ambienti per la sperimentazione in campo FI. FUSION prevede la creazione di un portale di scambio che permetta alle PMI di presentare le loro esigenze ed agli enti di ricerca di presentare i propri risultati, in maniera da venire incontro alle esigenze del mercato.

FIRE STATION vuole fornire alla rete dei progetti FIRE un nodo centrale che guida, coordina e armonizza domanda e offerta di *experimentation facilities* nel contesto della FI. A tal fine FIRE STATION ha creato un due organi appositi: il FIRE Office, che mette in contatto i soggetti che cercano risorse e ambienti

[230] URL: http://www.ict-fireworks.eu/

per la ricerca, e il FIRE Architecture Board, che invece coordina l'offerta di *experimentation facilities* affinché vengano elaborate una strategia e delle linee guida comuni. Il progetto FIRE STATION ha contribuito alla creazione di due portali online[231] , dove tali soggetti e in generale il pubblico interessato possano trovare notizie sulle iniziative in corso e sui progetti conclusi, insieme alle *road-map*, alle *call*, alle *best practices*, alle storie di successo nell'ambito della FI.

4.3.3 FIG – Future Internet Gateway

Il progetto qui esposto, FIG (Future Internet Gateway), ha in comune con altri progetti di CSA l'obiettivo di facilitare l'interazione e il reciproco scambio tra i soggetti coinvolti nella ricerca sulla FI, ma propone modalità più avanzate e innovative per raggiungere tale risultato. Il sistema FIG si presenta come una piattaforma online *user-friendly* e *self-learning* per la raccolta, la classificazione e la ricerca semantica di contenuti di argomento FI. Questi contenuti sono di diverso tipo e sono raccolti in maniera automatica e semi-automatica da fonti eterogenee. Si tratta dei risultati delle ricerche in ambito FI portate avanti sia dagli specifici progetti FIRE sia da altre iniziative o enti che non rientrano necessariamente nel programma di finanziamento europeo: libri digitali, paper scientifici, articoli di giornali o di riviste, *brochure* di eventi, pagine Web,ecc.

Tutto questo materiale è raccolto e organizzato internamente sulla base di una ontologia che è un sottoinsieme della vasta KB DBpedia. Il sistema ha un meccanismo di *ontology-based* Text Mining (il modulo TellMeFirst) che è in grado di estrarre dai documenti di testo i concetti che ne costituiscono l'argomento principale. Sulla base di tale meccanismo, i testi vengono classificati come appartenenti o meno al dominio della FI e sotto-classificati in un suo sottodomino (per es. "Cloud Computing" o "Smart Environment"). L' utente può esplorare il grafo concettuale che rappresenta il dominio della FI, scegliere l'argomento o gli argomenti che più gli interessano e visualizzare i contenuti classificati o connessi a quell'argomento da relazioni di somiglianza/contiguità. Può rendersi conto di quali enti di ricerca o singoli ricercatori sono coinvolti in quali campi di studio,

[231] URL: http://www.ict-fire.eu/home.html, http://www.future-internet.eu/

quali sono le tematiche più calde e quelle più di nicchia. La circoscrizione del dominio della FI partendo dalla KB multi-dominio di DBpedia avviene nella forma dell'autoapprendimento. Una volta raccolte le pubblicazioni ritenute più rilevanti nel settore della FI, esse vengono processate dal modulo TellMeFirst per estrarne una *bag of concepts* formata da un certo numero URI di DBpedia. Il sottoinsieme di DBpedia che comprende questi URI (con le classi a cui appartengono, le proprietà che possiedono e le relazioni che li interconnettono) è la Future Internet Knowledge Base (FIKB) della piattaforma.

La piattaforma FIG ha tra i suoi punti di forza anche la mantenibilità nel tempo. Prevede infatti un meccanismo di auto-alimentazione che sfrutta la classificazione semantica per valutare l'aderenza del materiale proveniente da diverse fonti rispetto agli argomenti affrontati dalla FI. Una volta agganciato a fonti di pubblicazione di nuovo materiale digitale, FIG è in grado di selezionare le pubblicazioni da classificare al proprio interno e quelle da scartare, senza l'intervento di specifiche figure professionali. L'archivio presente nel FIG è vivo e in continua espansione, arricchendosi e modellandosi sulla base delle evoluzioni del campo di conoscenza della FI. Il processo che va dall'acquisizione alla fruizione dei contenuti del portale FIG può essere suddiviso in quattro fasi, di seguito descritte.

4.3.3.1 Raccolta delle principali pubblicazioni in ambito Future Internet

La prima fase del progetto prevede l'intervento di esperti di dominio per la selezione delle pubblicazioni digitali più conosciute o accreditate nell'ambito della ricerca su Future Internet ed Internet Science. L'esperienza in questo settore del Centro Nexa su Internet & Società del Politecnico di Torino e della Fondazione AHREF[232] sono fondamentali per la riuscita delle operazioni di raccolta delle fonti. Si farà riferimento sia agli esiti dei progetti FIRE già conclusi e in corso (in particolare i portali http://future-internet.eu e http://ict-fire.eu, entrambi contributi del progetto FIRE STATION) sia alla rete tematica EINS sulla Internet Science. Il risultato sarà un insieme di documenti testuali in linguaggio naturale (inglese) da fornire in input al modulo di *topic extraction* di TellMeFirst. Il sof-

[232] URL: http://www.ahref.eu/it

tware produrrà in output una lista ordinata (*ranked*) di URI appartenenti al dominio di DBpedia.

4.3.3.2 Creazione della Future Internet Knowledge Base (FIKB)

Oggi è in costante aumento l'interesse per le KB come strumento per aumentare l'intelligenza dei motori di ricerca o per l'integrazione di dati e servizi sul Web. Ma la maggior parte delle basi di conoscenza coprono solo domini specifici, sono create da gruppi relativamente piccoli di ingegneri della conoscenza e sono molto costose da mantenere aggiornate parallelamente all'evoluzione del dominio. DBpedia, invece, sfruttando la fonte di conoscenza *cross-domain* e dinamica di Wikipedia, risulta essere sempre aggiornata e coprire un campo di conoscenze estremamente vasto. La base di conoscenza di DBpedia ha dunque diversi vantaggi rispetto alle altre KB esistenti:

1) copre numerosi settori;
2) rappresenta l'accordo semantico di una comunità di utenti di Internet;
3) si evolve automaticamente con le modifiche a Wikipedia;
4) è multilingue;
5) è accessibile sul Web in diversi formati (RDF, JSON, CSV).

Per ogni entità, DBpedia fornisce un identificatore univoco (URI) che può essere dereferenziato in base ai principi dei Linked Data. La semantica di ogni URI è definita dall'articolo di Wikipedia a cui l'URI è collegato. Un obiettivo fondamentale di FIG è quello di creare un'ontologia del dominio Future Internet, ovvero una mappa concettuale dei principali argomenti affrontati da questa ricerca. Tale operazione è di primaria importanza, in quanto senza di essa non si può avere un quadro preciso della dimensione del dominio, della sua stratificazione e profondità. Ogni tematica individuata all'interno della ricerca FI, sarà identificata in maniera univoca da una risorsa presente in DBpedia. Ciò consente di ottenere a priori un consenso semantico sulla particolare risorsa, senza dover negoziare o rinegoziare un insieme di significati. Il consenso semantico è lo stesso alla base delle voci presenti in Wikipedia, ottenuto in virtù dei suoi meccanismi di abilitazione dell'intelligenza collettiva online (vedi par. 1.2).

La FIKB è prodotta dal modulo di *topic extraction* di TellMeFirst ed è un sottoinsieme della KB di DBpedia. Essendo DBpedia estremamente ricca di connessioni sia al proprio interno sia verso l'esterno, tale ricchezza semantica è ereditata dalla FIKB e può essere sfruttata per operazioni avanzate di classificazione e ricerca.

4.3.3.3 Alimentazione del FIG

Il FIG conterrà una certa quantità di materiale inserito manualmente *una tantum* nel sistema, tra cui per esempio i documenti digitali utilizzati per la creazione della FIKB. Tuttavia il sistema sarà soprattutto un *gateway* continuamente alimentato in maniera automatica da risorse informative scelte sul Web. Le risorse saranno raccolte da diverse fonti . Tutti gli editori delle migliori riviste tecnico-scientifiche (Springer, Elsevier, Wiley, ACM, Oxford Journals, ecc.) offrono *feed* RSS per ricevere costantemente un flusso degli aggiornamenti contenente gli *abstract* e i metadati delle ultime pubblicazioni. Ma anche fonti meno tecniche come blog o quotidiani possono essere monitorate dal FIG. Il sistema si aggancia a queste fonti per ricevere costantemente nuovo materiale: ogni documento del flusso in ingresso viene preso in esame da TellMeFirst e categorizzato come interessante o meno in relazione alla ricerca FI. I documenti che passano la selezione del filtro "in o out" vengono in seguito classificati in sottocategorie specifiche appartenente al campo della FI, sulla base della FIKB. I documenti che non passano la selezione vengono semplicemente scartati.

4.3.3.4 Sottoclassificazione e ricerca semantica

I contenuti della piattaforma FIG sono classificati in maniera semantica in base ai concetti presenti nella FIKB. Per questo motivo il motore di ricerca che permette agli utilizzatori del sistema di accedere ai documenti può sfruttare le informazioni contenute nella KB per migliorare le sue *performance*. Sono numerosi i vantaggi offerti dall'utilizzo di una KB per fornire "intelligenza" a un motore di ricerca documentale (Hyvönen, 2012). I più tipici sono elencati di seguito.

1. Indicizzazione semantica. All'interno degli indici di ricerca, ogni dato è affiancato da specifiche informazioni semantiche, come l'appartenenza a

una o più classi (concetti), le relazioni che lo legano ad altri dati presenti nel dataset, o i sinonimi che lo identificano. In questo modo la ricerca si può basare sui concetti anziché sulle semplici *keyword*, migliorando la qualità dei risultati.

2. Analisi della query di ricerca. Particolari algoritmi riescono a ricondurre la query scritta dall'utente sulla maschera di ricerca a uno o più concetti presenti nell'ontologia. L'idea di base è quella di cercare nella base di conoscenza i percorsi logici che congiungono i concetti relativi alle parole digitate in linguaggio naturale, per scoprire il significato della richiesta al di là delle semplici parole chiave che la compongono (Lesmo et al., 2007).
3. Supporto multi linguistico. Grazie alla presenza nella knowledge base di sinonimi in varie lingue, il motore di ricerca è in grado di trovare correttamente documenti in lingua diversa da quella in cui è scritta la query dell'utente.
4. Navigazione semantica. Per mezzo della navigazione a faccette o dell'esplorazione di elementi grafici che rappresentano l'ontologia, è possibile interrogare il sistema senza digitare una query. In questo modo non soltanto si offre una modalità di fruizione del software più coinvolgente, ma si rende l'utente consapevole della ricchezza del dataset e della presenza di elementi che non si aspettava.
5. Suggerimenti di ricerca. Questi vengono forniti già in fase di digitazione, attraverso una funzione di auto-completamento del testo basata sulle entità presenti nella KB. Insieme ai risultati della ricerca, vengono forniti all'utente dei concetti correlati a cui potrebbe essere interessato.
6. Accesso alle informazioni in rete provenenti dai Linked Data *provider*. Questo significa un'espansione potenzialmente infinita della propria KB per mezzo del collegamento ad altri *repositories* in RDF presenti sul Web.

Conclusioni

Un acceso dibattito è nato in seguito alla pubblicazione nel 2010 sulla rivista Wired dell'articolo *The Web Is Dead: Long Live the Internet* di Chris Anderson e Michael Wolff. I due autori hanno paventato per la prima volta una crisi del Web: questo strumento, considerato fin dalla sua nascita come una delle più rivoluzionarie invenzioni della storia, sarebbe oggi minacciato da un sempre maggiore utilizzo della rete Internet attraverso protocolli diversi da HTTP. Negli ultimi anni, infatti, si è verificato un evidente spostamento degli utenti dal mondo aperto della Rete a piattaforme semichiuse, che utilizzano Internet per trasportare le informazioni, ma non il browser e il sistema ipertestuale per visualizzarle. Si parla per esempio delle *apps* per telefonia mobile, degli RSS, delle piattaforme di *instant messaging*, del VOIP, dei servizi di *streaming* multimediale, ecc.

> You wake up and check your email on your bedside iPad — that's one app. During breakfast you browse Facebook, Twitter, and The New York Times — three more apps. On the way to the office, you listen to a podcast on your smartphone. Another app. At work, you scroll through RSS feeds in a reader and have Skype and IM conversations. More apps. At the end of the day, you come home, make dinner while listening to Pandora, play some games on Xbox Live, and watch a movie on Netflix's streaming service. You've spent the day on the Internet — but not on the Web. (Anderson et al., 2010)

Il problema non riguarda solo aspetti tecnici come piattaforme e protocolli di rete, ma si estende ai contenuti e alla modalità in cui vengono distribuiti e fruiti. Nella maggior parte dei casi, le nuove applicazioni di Internet non seguono il modello di apertura e gratuità del Web, ma sono legate a modelli di business che prevedono la fruizione a pagamento di prodotti e servizi di una qualità percepita come "superiore" a quella del Web. I *provider* di questi prodotti e servizi mirano a convincere l'utente che esiste una Internet "di serie A" che deve necessariamente essere a pagamento, in quanto offre contenuti di maggior pregio o un più efficace controllo e filtraggio dei contenuti liberi. Al di fuori di questi «walled gardens» (giardini recintati) c'è invece la caotica messe dei contenuti Web, una

Internet “di serie B” dove la libertà si paga con la cattiva qualità dei contenuti, con i rischi per la sicurezza, con la carenza di privacy e, ovviamente, con il sovraccarico informativo.

Il Web è dunque di fronte a una sfida decisiva, in gioco c'è la sua stessa sopravvivenza. Cercare di porre rimedio ai difetti del Web visti nel paragrafo 4.1 e favorire al contrario la realizzazione di una nuova infrastruttura della conoscenza online è uno dei compiti fondamentali del Web Semantico e del movimento dei Linked Open Data. Progetti come DBpedia, che sfruttano le caratteristiche vincenti del Web già viste in Wikipedia (collaborazione, consenso, passione, dono) per creare un grande deposito di dati liberamente accessibili e interconnessi tra loro, aumentano la possibilità che il Web Semantico trovi una diffusione massiccia nella realtà di Internet e che venga utilizzato in contesti concreti. Queste tecnologie, che Jonathan Zittrain definisce «generative» (Zittrain, 2008), possono essere utilizzate in modo innovativo per la classificazione e la ricerca semantica dei documenti all'interno di uno spazio informativo.

Fare ordine sul Web categorizzandone formalmente le risorse è un beneficio soltanto se questa categorizzazione è scelta dagli utenti in maniera collaborativa, libera e consensuale. La classificazione deve ereditare ed esaltare la generatività presente in Wikipedia. Un mezzo o una tecnologia si dicono generative quando sono progettate per accettare qualsiasi contributo che segua un insieme basilare di regole (Zittrain, 2008, p. 3). Stabilite le regole del gioco (della classificazione dei documenti, sia manuale che automatica, così come della scrittura delle voci di Wikipedia), il processo successivo seguirà le logiche informali, spesso inspiegabili ma vincenti, dell'espansione del Web.

Questa tesi, per quanto ambiziosa e multidisciplinare, vuole solo dare un modesto contributo a quella che in futuro potrebbe essere una prassi comune. Le tecnologie di Natural Language Processing, di Machine Learning e di Text Mining per la classificazione e la ricerca semantica dei documenti sono oggi utilizzate dalle grandi aziende per controllare, monitorare e comprendere la propria clientela, ma possono anche servire agli utenti di Internet per organizzare il proprio universo cognitivo, senza alcun condizionamento esterno e proprietario. Gli algoritmi utilizzati da TellMeFirst, così come quelli alla base di altri progetti *o-*

pen source nello stesso ambito, sono liberamente accessibili per lo studio, l'utilizzo, la modifica e il miglioramento[233]. Si può sfuggire alla logica del PageRank di Google solo se il dislivello di consapevolezza tecnologica tra le grandi Aziende e le Università che finanziano la ricerca *open* rimane colmabile. E questo può essere raggiunto soprattutto garantendo ai ricercatori e agli sviluppatori *open source* un livello economico e di gratificazione personale competitivo con le corrispondenti condizioni aziendali. Senza uno sforzo economico da parte di governi e istituzioni, il pericolo è che, in un futuro più o meno prossimo, il Web concluda la sua corsa come le altre ideologie egualitarie della storia, che, nonostante gli esordi più promettenti, hanno presto conosciuto un lento ma inesorabile declino.

[233] URL: https://github.com/TellMeFirst/tellmefirst

Bibliografia

Alesso, H. P., & Smith, C. F. (2009). *Thinking on the Web: Berners-Lee, Gödel and Turing*. Hoboken: John Wiley & Sons.

Allemang, D., & Hendler, J. A. (2008). *Semantic Web for the working ontologist: effective modeling in RDFS and OWL*. San Francisco, Calif.: Morgan Kaufmann.

Alpert, J. & Hajaj, N. (2008). *We knew the Web was big...* (URL: http://googleblog.blogspot.it/2008/07/we-knew-web-was-big.html).

Anderson, C. & Wolff, M. (2010). *The Web Is Dead. Long Live the Internet* (URL: http://www.wired.com/magazine/2010/08/ff_webrip/).

Anderson, T. (2008). *The problem with Wikipedia and bias.* (URL: http://www.onlineopinion.com.au/view.asp?article=6954).

Ayers, P., Matthews, C., & Yates, B. (2008). *How Wikipedia works: and how you can be a part of it*. San Francisco: No Starch Press.

Basili, R., & Moschitti, A. (2005). *Automatic text categorization: from information retrieval to support vector learning*. Roma: Aracne.

Benkler, Y. (2002). Coase ' s Penguin , or , Linux and The Nature of the Firm. *Yale Law Journal*, *112*(3), 369-446. Yale University Press.

Benkler, Y. (2006). *The wealth of networks: how social production transforms markets and freedom*. New Haven, Conn.: Yale University Press.

Berners-Lee, T. (2009). *Putting Government Data online.* (URL: http://www.w3.org/DesignIssues/GovData.html).

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. (K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, et al., Eds.)*Scientific American*, *284*(5), 34-43. Citeseer.

Berra, M., & Meo, A. R. (2006). *Libertà di software, hardware e conoscenza: informatica solidale 2*. Torino: Bollati Boringhieri.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. (T Heath, M. Hepp, & C Bizer, Eds.)*International Journal on Semantic Web and Information Systems*, *5*(3), 1-22. Elsevier.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - A crystallization point for the Web of Data. *Web Semantics Science Services and Agents on the World Wide Web*, 7(3), 154-165. Elsevier.

Blair, A. (2010). *Too much to know: managing scholarly information before the modern age*. New Haven, Conn.: Yale University Press.

Cairo, F., & Futia, G. (2011). *ICONVIS: an Interactive and Customizable System for Semantic Data Visualization*, Poster Session, ASSYST–FuturICT International.

Cairo, F., & Muraca, A. (2012). Il motore semantico Cultura per la valorizzazione del patrimonio digitale piemontese. In Monaci S., Ilardi E., Spano M. (Eds.), *Patrimonio virtuale. Modelli di comunicazione e tecnologie per la valorizzazione dei beni culturali*. Napoli:ScriptaWeb.

Carr, N. G. (2010). *The shallows: what the Internet is doing to our brains*. New York: W.W. Norton.

Cassell, J. (2011). *Editing Wars Behind the Scenes.* (URL: http://www.nytimes.com/roomfordebate/2011/02/02/where-are-the-women-in-wikipedia/a-culture-of-editing-wars).

Castells, M. (2001). *The Internet galaxy: reflections on the Internet, business, and society*. Oxford: Oxford University Press.

Cheng, W., & Hüllermeier, E. (2009). Combining instance-based learning and logistic regression for multilabel classification. (W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor, Eds.)*Machine Learning*, *76*(2-3), 211-225. Springer.

Cohen, M. (2008). *Encyclopaedia Idiotica*. (URL: http://www.cautbulletin.ca/en_article.asp?articleid=2693).

Cohen, N. (2011). *Define Gender Gap? Look Up Wikipedia's Contributor List* (URL: http://www.nytimes.com/2011/01/31/business/media/31link.html ?_r=0).

Csikszentmihalyi, M. (1990). *Flow: the psychology of optimal experience*. New York: Harper & Row.

Denning, P., Horning, J., Parnas, D., & Weinstein, L. (2005). Wikipedia risks.*Communications of the ACM*, *48*(12), 152. ACM.

Donato, F. (2010). *Lo Stato trasparente: linked open data e cittadinanza attiva*. Pisa: ETS.

Durkheim, E. (1912). *Les formes élémentaires de la vie religieus*. Paris:Alcan (Durkheim, E., *Le forme elementari della vita religiosa*, Roma: Maltemi editore, 2005).

Eco, U. (1995). *A Conversation on Information*. (URL: http://carbon.ucdenver.edu/~mryder/itc/eco/eco.html).

Eco, U. (2012). *Che casino, troppe informazioni*. (URL: http://espresso.repubblica.it/dettaglio/che-casino-troppe-informazioni/2189205).

Ferragina, P., & Scaiella, U. (2010). TAGME : On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). *Proceedings of the 19th ACM*

international conference on Information and knowledge management, CIKM '10, 1625-1628. ACM.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. (F. R. Palmer, Ed.)*Studies in Linguistic Analysis*, *1952-59*, 1–31. Longman.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*(2), 171-191. American Psychological Association.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time.*American Psychologist*, *54*(1), 5-20. American Psychological Association.

Futia, G. (2011). *Linked Data ICONVIS. Progetto e sviluppo di un visualizzatore di dati relazionali basato su ontologie*. Master's Degree Thesis. Politecnico di Torino.

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the Brittleness Bottleneck using Wikipedia : Enhancing Text Categorization with Encyclopedic Knowledge. *Proceedings Of The National Conference On Artificial Intelligence*,*21*(2), 1301-1306. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L. (2002). Sweetening Ontologies with DOLCE. (A. Gómez-Pérez & V. R. Benjamins, Eds.)*Lecture Notes in Computer Science*, *2473*, 166-181. Springer.

Giuliano, C., Gliozzo, A. M., & Strapparava, C. (2009). Kernel Methods for Minimally Supervised WSD. *Computational Linguistics*, 35(4), 513-528.

Goth, G. (2010). The Global Net Neutrality Debate: Back to Square One? *IEEE Internet Computing.*IEEE.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. (H. Burkhardt & B. Smith, Eds.)*Knowledge Creation Diffusion Utilization*,*5*(April), 199-220. Citeseer.

Guarino, N., & Welty, C. (2002). Evaluating ontological decisions with OntoClean. *Communications of the ACM*, *45*(2), 61-65. ACM.

Gupta, M., Yin, Z., Han, J., & Li, R. (2010). Survey on Social Tagging Techniques University of Techniques. *SIGKDD Explorations*, *12*(1), 58-72. ACM.

Haarslev, V., & Möller, R. (2001). RACER System Description. (R. Goré, A. Leitsch, & T. Nipkow, Eds.)*Syntax And Semantics*, *1*(2083), 701-705. Springer.

Han, X., & Sun L. (2011). A Generative Entitymention Model for Linking Entities with Knowledge Base. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 945–954, Association for Computational Linguistics.

Harding, T. D., & Tait, J. (2002). *64 great chess games: instructive classics from the world of correspondence chess*. Dublin: Chess Mail.

Harris, Z. (1954). Distributional structure. (J. A. Fodor & J. J. Katz, Eds.) *Word Journal Of The International Linguistic Association*, *10*(23), 146-162. Oxford University Press.

Hart, G., & Dolbear, C. (2013). *Linked data: a geographic perspective*. Boca Raton: Taylor & Francis.

Huang, Z., Chen, H., Yu, T., Sheng, H., Luo, Z., & Mao, Y. (2009). *Semantic Text Mining with Linked Data*. 2009 Fifth International Joint Conference on INC IMS and IDC, 338-343. Ieee.

Hyvönen, E. (2012). *Publishing and using cultural heritage linked data on the semantic Web*. San Rafael, Calif.: Morgan & Claypool Publishers.

Ingersoll, G. S., & Morton, T. S. (2013). *Taming text: how to find, organize, and manipulate it*. Shelter Island, NY: Manning.

Jenkins, H. (2006[1]). *Convergence culture: where old and new media collide*. New York: New York University Press.

Jenkins, H. (2006[2]). *Fans, bloggers, and gamers: exploring participatory culture*. New York: New York University Press.

Ji, H., Grishman, R., & Dang, H. (2011). Overview of the TAC2011 Knowledge Base Population Track. *Proceedings of the Text Analysis Conference* (TAC 2011).

Johnson, C. A. (2012). *The information diet: a case for conscious consumption*. Beijing: O'Reilly Media.

Klingberg, T. (2009). *The overflowing brain: information overload and the limits of working memory*. Oxford: Oxford University Press.

Kovacs, B. (1990). *The decision-making process for library collections: case studies in four types of libraries*. New York: Greenwood Press.

Kudelka, M., Snasel, V., El-Qawasmeh, E., Lehecka, O., & Tesarik, J. (2007). Domain Patterns and Semantic Annotation of Web Pages. *2006 1st International Conference on Digital Information Management*, 504-510. IEEE.

Kujoth, J. S. (1969). *Libraries, readers, and book selection*. Metuchen, N.J.: Scarecrow Press.

Lesmo, L., & Robaldo, L. (2007). Use of ontologies in Practical NL Query Interpretation. *Lecture Notes In Artifcial Intelligence* (LNAI, vol.4733), Springer.

Lévy, P. (1994). *L'Intelligence collective: pour une anthropologie du cyberspace*. Paris: La Découverte (Lévy, P., *L'intelligenza collettiva: per un'antropologia del cyberspazio*. Milano: Feltrinelli, 1996).

Lih, A. (2009). *The Wikipedia revolution: how a bunch of nobodies created the world's greatest encyclopedia*. New York: Hyperion.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

Marrero, M., Urbano, J., Morato, J., & Sánchez-Cuadrado, S. (2010). *On the Definition of Patterns for Semantic Annotation*. Proceedings of the third workshop on Exploiting semantic annotations in information retrieval (pp. 15-16). ACM.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., & Oltramari, A. (2003). WonderWeb Deliverable D18. (URL: http://www.loa.istc.cnr.it/Papers/DOLCE2.1-FOL.pdf).

Maurer, H., Balke, T., Kappe, F., Kulathuramaiyer, N., Weber, S. & Zaka, B. (2007). *Report on dangers and opportunities posed by large search engines, particularly Google* . Technical report for Austrian Federal Ministry of Transport.

McCarthy, D. (2009). Word Sense Disambiguation: An Overview. *Language and Linguistics Compass*, *18*(2), 537-558. Wiley.

McHernry , R. (2004). *The Faith-Based Encyclopedia* (URL: http://www.ideasinactiontv.com/tcs_daily/2004/11/the-faith-based-encyclopedia.html).

Meeting on Visualization in Complex Environments, Turin.

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia Spotlight : Shedding Light on the Web of Documents. *Text*, *95*(2), 1-8. Facultad de Informática (UPM).

Mendes, P.N., Jakob, M., Bizer, C. (2012). DBpedia for NLP: A Multilingual Cross-domain Knowledge Base. *Proceedings of the International Conference on Language Resources and Evaluation*, LREC 2012, 21-27 May 2012, Istanbul.

Mihalcea, R. (2007). *Using Wikipedia for Automatic Word Sense Disambiguation*. English, 2007(April), 196-203. Association for Computational Linguistics.

Mihalcea, R., & Csomai, A. (2007). *Wikify!: linking documents to encyclopedic knowledge*. In M. J. Silva, A. H.F. Laender, R. A. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, & A. O. Falcão (Eds.), CIKM 07 Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (pp. 233-242). ACM.

Milicevic, A. K., Nanopoulos, A., & Ivanovic, M. (2010). Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions. Artificial Intelligence Review, 33(3), 187-209. Springer Netherlands.

Monaci, S. (2008). *La conoscenza on line: logiche e strumenti*. Roma: Carocci.

Morozov, E. (2011). *The net delusion: the dark side of Internet freedom*. New York, NY: PublicAffairs.

Muñoz-García, O., García-Silva, A., Corcho, Ó., De La Higuera Hernández, M., & Navarro, C. (2011). Identifying Topics in Social Media Posts using DBpedia. (J.-D. Meunier, H. Hrasnica, & F. Genoux, Eds.)*Media*, *28020*, 81-86. Facultad de Informática (UPM).

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, *41*(2), 1-69. ACM.

Nielsen, M. A. (2012). *Reinventing discovery: the new era of networked science*. Princeton, N.J.: Princeton University Press.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Internet And Web Information Systems*, *54*(2), 1-17. Technical report, Stanford Digital Library Technologies Project, 1998.

Pariser, E. (2011). *The filter bubble: what the Internet is hiding from you.* New York: Penguin Press.

Pease, A., Niles, I., & Li, J. (2002). The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. *Imagine* (Vol. 28, pp. 7-10). AAAI Press.

Ratinov, L., Roth, D., Downey, D., & Anderson, M. (2011). Local and Global Algorithms for Disambiguation to Wikipedia. *Computational Linguistics*, 1, 1375-1384. Association for Computational Linguistics.

Raymond, E. S. (1999). *The cathedral & the bazaar: musings on Linux and open source by an accidental revolutionary.* Beijing: O'Reilly.

Reagle, J. M. (2010). *Good faith collaboration: the culture of Wikipedia.* Cambridge, Mass.: MIT Press.

Rheingold, H. (2002). *Smart mobs: the next social revolution.* Cambridge, MA: Perseus Pub.

Rosenzweig, R. (2006). Can History Be Open Source? Wikipedia and the Future of the Past. *Journal of American History*, *93*(1), 117-146. Oxford University Press.

Runggaldier, E., & Kanzian, C. (1998). *Grundprobleme der analytischen Ontologie.* Paderborn: F. Schöningh. (Runggaldier, E., & Kanzian, C., *Problemi fondamentali dell'ontologia analitica.* Milano: Vita e pensiero, 2002).

Rusu, D., Fortuna, B. and Mladenic, D. 2011. Automatically Annotating Text with Linked Open Data. *4th Linked Data on the Web Workshop* (LDOW 2011), 20th World Wide Web Conference (WWW 2011). Hyderabad, India.

Ryan, J. (2010). *A history of the Internet and the digital future*. London, England: Reaktion Books.

Sammut, C., & Webb, G. I. (2010). *Encyclopedia of machine learning*. New York: Springer.

Sanger, L. (2004). *Why Wikipedia Must Jettison Its Anti-Elitism.* (URL: http://www.kuro5hin.org/story/2004/12/30/142458/25).

Sapolsky, R. M. (1994). *Why zebras don't get ulcers: a guide to stress, stress related diseases, and coping*. New York: W.H. Freeman.

Scarborough, R. (2010). *Wikipedia Whacks the Right.* (URL: http://www.humanevents.com/2010/09/27/wikipedia-whacks-the-right/).

Schrage, M. (1990). *Shared minds: the new technologies of collaboration*. New York: Random House.

Schreiber, G. (2000). *Knowledge engineering and management the CommonKADS methodology*. Cambridge, Mass.: MIT Press.

Shirky, C. (2008). *Here comes everybody: the power of organizing without organizations*. New York: Penguin Press.

Sirin, E., Parsia, B., Grau, B., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Web Semantics Science Services and Agents on the World Wide Web*, *5*(2), 51-53. Elsevier.

Spoerri, A. (2007). What is popular on Wikipedia and why? *First Monday*, *12*(4), 1-22.

Stokes, S. (2009). *Digital copyright: law and practice* (3rd ed.). Oxford: Hart Pub.

Sunstein, C. R. (2006). *Infotopia: how many minds produce knowledge*. Oxford: Oxford University Press.

Surowiecki, J. (2004). *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Doubleday.

Toffler, A. (1970). *Future shock*. New York: Random House.

Trant, J. (2009). Studying Social Tagging and Folksonomy : A Review and Framework. *Journal Of Digital Information*, *10*(1), 1-42.

Weinberger, D. (2011). *Too big to know: rethinking knowledge now that the facts aren't the facts, experts are everywhere, and the smartest person in the room is the room*. New York: Basic Books.

Wheeler, W. M. (1911), The ant-colony as an organism. *J. Morphol.*, 22: 307–325.

Wilson, E. J. (2005). What is Internet Governance and Where Does it Come From*?* *Journal of Public Policy*, 25(1), 29-50. Cambridge University Press.

Yang, Y. (2001). A Study on Thresholding Strategies for Text Categorization. (W. B. Croft, D. J. Harper, D. H. Kraft, & J. Zobel, Eds.) *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 01*, (9), 137-145. ACM Press.